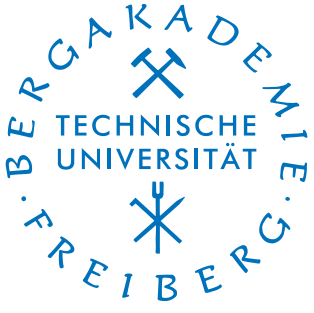
Rational Krylov Methods for Operator Functions

Von der Fakultät für Mathematik und Informatik
der Technischen Universität Bergakademie Freiberg
genehmigte Dissertation zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.),
vorgelegt von Dipl.-Math. Stefan Güttel,
geboren am 27.11.1981 in Dresden.



Betreuer:	Prof. Dr. Michael Eiermann (Freiberg)
Gutachter:	Prof. Dr. Axel Ruhe (Stockholm)
	Prof. Dr. Nick Trefethen FRS (Oxford)
Verleihung:	Freiberg, am 12.03.2010

This page is intentionally left almost blank.

Abstract

We present a unified and self-contained treatment of rational Krylov methods for approximating the product of a function of a linear operator with a vector. With the help of general rational Krylov decompositions we reveal the connections between seemingly different approximation methods, such as the Rayleigh–Ritz or shift-and-invert method, and derive new methods, for example a restarted rational Krylov method and a related method based on rational interpolation in prescribed nodes. Various theorems known for polynomial Krylov spaces are generalized to the rational Krylov case. Computational issues, such as the computation of so-called matrix Rayleigh quotients or parallel variants of rational Arnoldi algorithms, are discussed. We also present novel estimates for the error arising from inexact linear system solves and the approximation error of the Rayleigh–Ritz method. Rational Krylov methods involve several parameters and we discuss their optimal choice by considering the underlying rational approximation problems. In particular, we present different classes of optimal parameters and collect formulas for the associated convergence rates. Often the parameters leading to best convergence rates are not optimal in terms of computation time required by the resulting rational Krylov method. We explain this observation and present new approaches for computing parameters that are preferable for computations. We give a heuristic explanation of superlinear convergence effects observed with the Rayleigh–Ritz method, utilizing a new theory of the convergence of rational Ritz values. All theoretical results are tested and illustrated by numerical examples. Numerous links to the historical and recent literature are included.

Acknowledgements

I thank Prof. Dr. Michael Eiermann who made this thesis possible in the first place. Already when I was an undergraduate student he directed my interest towards Krylov methods and potential theory. I appreciate his constant support and belief in me, and the inspiring mathematical discussions.

I am grateful to Prof. Dr. Oliver Ernst for his help. In particular, I thank him for reading all the drafts and suggesting countless improvements and corrections regarding content and language. Of course, I take full responsibility for remaining errors and typos in this thesis.

I thank my colleagues and friends at TU Bergakademie Freiberg, in particular at the Institutes of Numerical Mathematics & Optimization and Geophysics. I learned a lot in the past three years of fruitful collaboration with them.

I thank Prof. Dr. Bernhard Beckermann for his hospitality during an intensive research week in Lille in February 2009, the outcome of which was a joint paper with Prof. Dr. Raf Vandebril, to whom I am also grateful. I thank Dr. Maxim Derevyagin for discussions about linear operators, and Prof. Dr. Mike Botchev for translating parts of the paper by A. N. Krylov. In 2008 I enjoyed two months of Japanese summer, working at the National Institute of Informatics, and I am thankful to Prof. Dr. Ken Hayami for this unforgettable experience. I thank Prof. Dr. Axel Ruhe and Prof. Dr. Nick Trefethen for serving on the reading committee. I acknowledge financial support by the Deutsche Forschungsgemeinschaft.

Most importantly, I thank my parents for their invaluable support.

Contents

List of Figures

Notation

By $\subseteq$ we denote set inclusion with possible equality, whereas $\subset$ denotes strict inclusion.

Unless stated otherwise, the following conventions for variable names hold: Small Latin letters $(a, b, \ldots)$ and Greek letters $(\alpha, \beta, \Phi, \Psi, \ldots)$ denote scalars or scalar-valued functions (including measures $\mu, \nu, \ldots$). Bold Latin letters $(\boldsymbol{a}, \boldsymbol{b}, \ldots)$ stand for vectors, and these are usually in column format. Capital Latin letters $(A, B, \ldots)$ stand for linear operators. Calligraphic letters $(\mathcal{P}, \mathcal{Q}, \ldots)$ denote linear spaces and double-struck capital letters $(\mathbb{D}, \mathbb{R}, \ldots)$ are special subsets of the complex plane $\mathbb{C}$.

Here is a list of frequently used symbols with the page numbers of their first occurrence in this thesis.

1 Introduction

Probable impossibilities
are to be preferred to
improbable possibilities.
Aristotle

An important problem arising in science and engineering is the computation of

$$f(A)\boldsymbol{b},$$

where A is a linear operator, $\boldsymbol{b}$ is a vector, and f is a function such that $f(A)$ is defined. In this thesis we consider approximations for $f(A)\boldsymbol{b}$, denoted by $\boldsymbol{f}_m$, which can be represented as

$$\boldsymbol{f}_m = r_m(A)\boldsymbol{b},$$

where r_m is a rational function of type $(m-1, m-1)$ with a prescribed denominator. Often the function $f = f^\tau$ depends on a parameter τ, and consequently the same is true for the corresponding approximations $\boldsymbol{f}_m^\tau$. Here is a selection of functions that are of particular interest for applications:

— the resolvent or transfer function $f^\tau(z) = (z - \tau)^{-1}$, arising in model reduction problems in the frequency domain [Ruh84, GGV96],

— the exponential function $f^\tau(z) = \exp(\tau z)$ or variants of it, for the solution of evolution problems [HLS98, ST07b],

— variants of trigonometric functions such as $f^\tau(z) = \operatorname{sinc}(\tau\sqrt{z})$, for the solution of time-dependent hyperbolic problems [GH08],

— fractional powers $f(z) = z^\alpha$, in particular with $\alpha = 1/2$ for the solution of stochastic differential equations in population dynamics [All99] and neutron transport [SA00, ABB00], or for preconditioning domain decomposition methods [AL08],

— the sign function $f(z) = \operatorname{sgn}(z)$, arising in quantum chromodynamics [EFL+02].

In this thesis we consider *rational Krylov methods* for computing the approximations $\boldsymbol{f}_m^\tau$. These methods generalize standard (polynomial) Krylov methods for the approximation of operator functions since polynomials can be interpreted as rational functions with all poles at infinity. It is useful to view rational (and polynomial) Krylov methods as abstract *approximation methods*, which can be characterized by two components:

(a) an m-dimensional *search space* $\mathcal{V}_m$ from which the approximations $\boldsymbol{f}_m^\tau$ are chosen, and

(b) the *extraction*, i.e., *how* the approximations $\boldsymbol{f}_m^\tau$ are chosen from $\mathcal{V}_m$.

We now use a simple model problem to briefly describe and compare two approximation methods, a polynomial and a rational Krylov method.

1.1 A Model Problem

Consider the initial-boundary value problem for the heat equation on the unit cube in d dimensions

$$\partial_\tau u = \Delta u \qquad \text{in } \Omega = (0,1)^d,\ \tau > 0, \tag{1.1a}$$

$$u(\boldsymbol{x}, \tau) = 0 \qquad \text{on } \Gamma = \partial\Omega,\ \tau > 0, \tag{1.1b}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \qquad \text{in } \Omega, \tag{1.1c}$$

which is a standard model problem in the literature [GS92, EH06].

When the Laplacian is discretized by the usual $(2d+1)$-point stencil on a uniform grid involving n interior grid points in each Cartesian direction, problem (1.1) reduces to the

initial value problem

$$\boldsymbol{u}'(\tau) = A\boldsymbol{u}(\tau), \quad \tau > 0, \tag{1.2a}$$

$$\boldsymbol{u}(0) = \boldsymbol{b}, \tag{1.2b}$$

with a matrix $A \in \mathbb{R}^{N\times N}$ ($N = n^d$) and an initial vector $\boldsymbol{b} \in \mathbb{R}^{N\times N}$ consisting of the values $u_0(\boldsymbol{x})$ at the grid points $\boldsymbol{x}$. The solution of (1.2) is given by

$$\boldsymbol{u}(\tau) = f^\tau(A)\boldsymbol{b}, \quad \text{where } f^\tau(z) = \exp(\tau z). \tag{1.3}$$

It is known that A is symmetric and its eigenvalues $\Lambda(A)$ are contained in the interval $(-4d(n+1)^2, 0)$. Note that A becomes large as we increase the number n, but it remains sparse since there are at most $2d+1$ nonzeros per row. Here are two well-known approximation methods for computing approximations $\boldsymbol{f}_m^\tau$ for (1.3).

The Polynomial Lanczos Method. This method utilizes an orthonormal basis $V_m = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m] \in \mathbb{C}^{N\times m}$ of the polynomial Krylov space

$$\mathcal{K}_m(A, \boldsymbol{b}) := \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}\},$$

satisfying a Lanczos decomposition

$$AV_m = V_m T_m + \boldsymbol{v}_{m+1}\beta_m \boldsymbol{e}_m^T,$$

where $\boldsymbol{v}_{m+1} \perp \mathcal{R}(V_m)$ and T_m is a symmetric tridiagonal matrix [DK89, GS92]. The approximations are then computed as

$$\boldsymbol{f}_m^\tau := V_m f^\tau(T_m) V_m^* \boldsymbol{b}.$$

Hence, coming back to our two components of approximation methods, the search space is $\mathcal{V}_m = \mathcal{K}_m(A, \boldsymbol{b})$ and the extraction is called *Rayleigh extraction* (because $T_m = V_m^* A V_m$ is a matrix Rayleigh quotient of A).

The Shift-and-Invert Lanczos Method. This method utilizes an orthonormal basis V_m of a rational Krylov space $\mathcal{V}_m = \mathcal{K}_m((A-\xi I)^{-1}, \boldsymbol{b})$ for some shift $\xi \in \mathbb{C} \setminus \Lambda(A)$ satisfying

$$(A-\xi I)^{-1} V_m = V_m T_m + \boldsymbol{v}_{m+1}\beta_m \boldsymbol{e}_m^T,$$

which is also a Lanczos decomposition, but now for the shifted and inverted operator $(A-\xi I)^{-1}$ [MN04, EH06]. The approximations are then extracted by back-transforming the matrix T_m, i.e.,

$$\boldsymbol{f}_m^\tau := V_m f^\tau(T_m^{-1} + \xi I_m) V_m^* \boldsymbol{b}.$$

Comparison of the Two Approximation Methods. In Figure 1.1 we compare the convergence of the polynomial and the rational Krylov method for the model problem in $d = 3$ dimensions and a fixed time parameter $\tau = 0.1$. In the three plots we vary the grid parameter $n \in \{15, 31, 63\}$. The linear systems in the shift-and-invert method were solved by a geometric multigrid method to a relative residual norm of 10^{-14} and we used the (rather arbitrary) shift $\xi = 1$. We observe that the number of iterations m required by the polynomial Krylov method to reach a certain accuracy is roughly proportional to n. Indeed, it is proven in [HL97] that for negative definite operators the regime of superlinear convergence is reached after $\sqrt{\|\tau A\|} = O(n)$ iterations. The costly part of each polynomial Krylov iteration is essentially a matrix-vector product with A, which can be performed in $O(N)$ operations, and two orthogonalizations at the cost of $O(N)$ operations to build the orthonormal Krylov basis V_m. On the other hand, the number of iterations required by the rational Krylov method appears to be independent of n. Under the assumption that each iteration involves a linear system of size $N \times N$ that can be solved in $O(N)$ operations, the overall cost for solving a fixed number of these systems is $O(N)$. Additionally, one evaluation of a function of T_m typically involves $O(m^2)$ or even $O(m^3)$ operations and this cost becomes non-negligible for the polynomial Krylov method as m gets large.

	Polynomial Lanczos	Shift-and-invert
Iterations	$m = O(n)$	$m = \text{const.}$
Operator/Iteration	$O(N)$	$O(N)$
Orthogonalize/Iteration	$O(N)$	$O(N)$
Evaluate $f(T_m)$	$O(m^2)$	const.
Total	$O(nN)$	$O(N)$

Table 1.1: Operation counts ($N = n^d$)

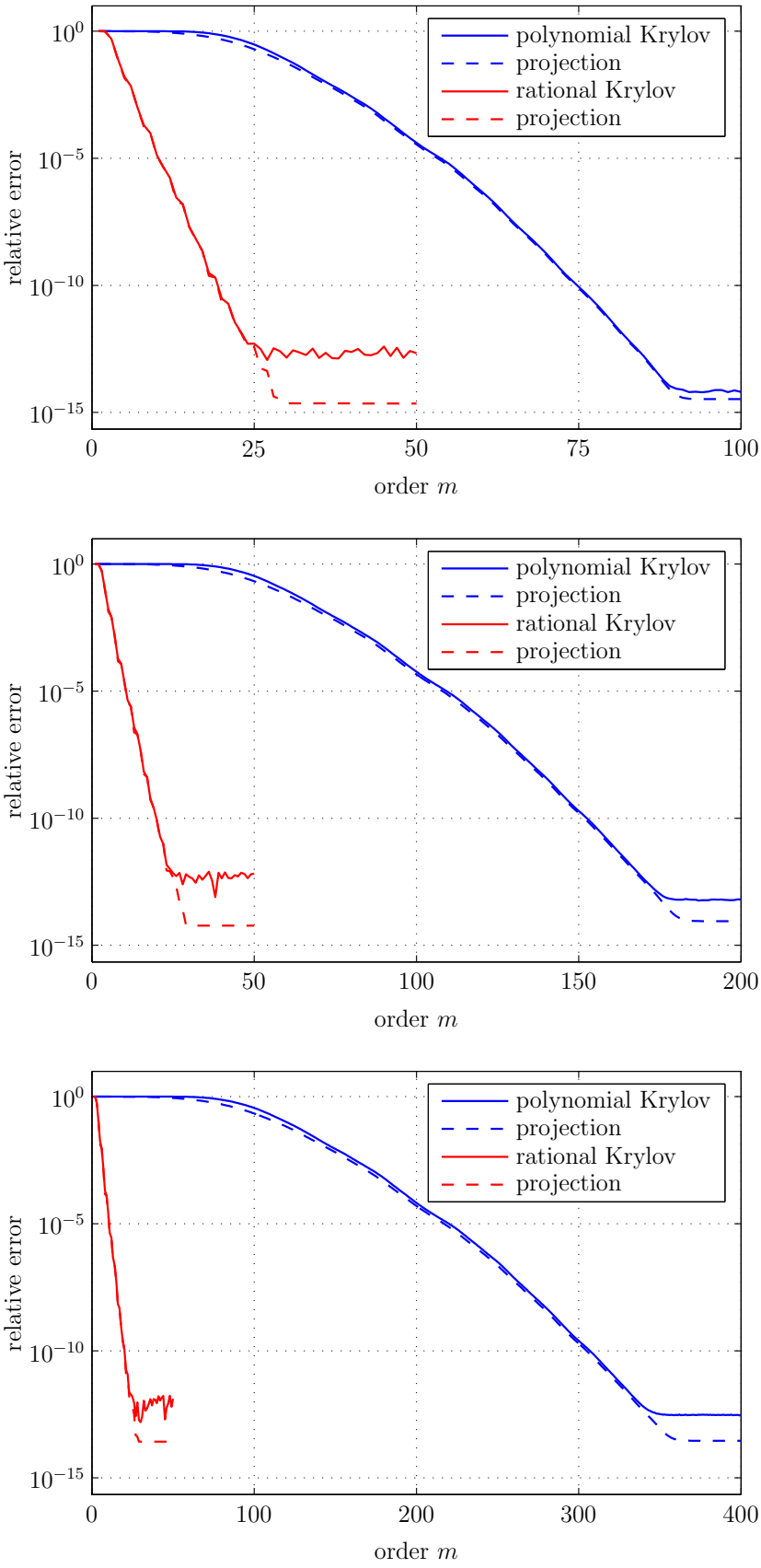


Figure 1.1: Convergence curves of a polynomial and a rational Krylov method for the 3D heat equation discretized with $n \in \{15, 31, 63\}$ grid points in each coordinate direction, respectively. The dashed line shows the error of the orthogonal projection of the exact solution $f^{\tau}(A)\boldsymbol{b}$ onto the respective Krylov spaces.

The total operation counts summarized in Table 1.1 on page 4 indicate that for a large enough grid parameter n the rational Krylov method will ultimately outperform the polynomial Krylov method under the assumption that we employ a linear system solver that is (asymptotically) as expensive as a matrix-vector product with A. This becomes even more pronounced when A is not self-adjoint so that the cost for orthogonalization grows in each iteration, though, on the other hand, the linear systems may then be more difficult to solve. To say it in the words of Aristotle, it is therefore not an "improbable possibility" that rational Krylov methods can be useful in practice. Some of the reasons why rational Krylov methods are not (yet) as popular as polynomial Krylov methods are:

(−) The implementation of rational Krylov methods is more involved than that of polynomial methods, e.g., we need to solve shifted linear systems with A.

(−) The cost of the linear system solves only pays off for very large problems (if ever).

(−) Rational Krylov methods are not parameter-free, e.g., in our example we could possibly find a better shift than $\xi = 1$.

(−) The convergence behavior is not completely understood, in particular in the presence of inexact solves of the linear systems.

On the other hand, rational Krylov methods are powerful:

(+) Rational approximation may be more efficient than polynomial approximation. To give an example, we consider the exponential function $f(x) = \exp(x)$ on the negative real axis $(-\infty, 0]$. The polynomial with the lowest uniform error 0.5 is obviously $p(x) \equiv 0.5$, whereas even a simple rational function such as $r(x) = (1 - x/m)^{-m}$ can achieve an arbitrarily small uniform error on $(-\infty, 0]$ for sufficiently large m.

(+) Rational functions possess a partial fraction expansion, which yields an immediate and effective method for parallelizing rational Krylov methods (though stability problems may arise).

Last but not least, rational Krylov methods are interesting because there are obviously many open questions connected to them, some of which we hope to answer in this thesis.

1.2 Aim and Structure of this Thesis

Our aim is to develop and study rational Krylov methods for the approximation of $f(A)\boldsymbol{b}$ in a unified way, to treat different aspects of their implementation, and to test them on numerical examples.

To make this thesis self-contained, Chapter 2 is devoted to the definition of functions of linear operators A. Although in numerical computations A is usually a matrix, it often originates from the discretization of a bounded or unbounded linear operator. A main advantage of rational Krylov methods is that their convergence can be independent of the discretization mesh width (as in the above model problem) and therefore it is adequate to study these methods more generally for linear operators.

In Chapter 3 we introduce the Rayleigh method as the canonical extraction associated with a search space $\mathcal{V}_m$. When $\mathcal{V}_m$ is generated by polynomials or rational functions in A times $\boldsymbol{b}$, the Rayleigh method becomes the Rayleigh–Ritz method and possesses interesting properties such as an interpolation characterization and near-optimality. We study these properties in Chapter 4 without utilizing any Krylov decomposition for $\mathcal{V}_m$. Such rational Krylov decompositions enter the stage in Chapter 5, where we also discuss the rational Arnoldi algorithm invented by A. Ruhe.

Chapter 6 deals with several computational issues, e.g., we investigate inexact solves, multigrid methods for shifted linear systems, and make comments about the parallel implementation of the rational Arnoldi algorithm. In Chapter 7 we study selected rational approximation problems, which are naturally associated with rational Krylov methods. In Chapter 8 we give insights into the interpolation procedure underlying the Rayleigh–Ritz method by discussing the convergence of rational Ritz values. In Chapter 9 we test and illustrate the theoretical results with the help of numerical experiments.

Each chapter begins with a brief overview, where we often review contributions from the historical literature. Whenever we cite many references in a row, such as [Vor87, DK89, Kni91], this is a possibly non-exhaustive list and should be read as "see, e.g., [Vor87, DK89, Kni91] and the references therein." Within chapters we use common counters for Definitions, Lemmas, Remarks, etc. We believe this simplifies referencing for the reader since finding **Remark 1.3** tells one that **Lemma 1.4** should follow.

The computations in this thesis were carried out in MATHWORKS MATLAB (release 2008b), partly using the *Schwarz–Christoffel toolbox* by T. A. Driscoll [Dri96, Dri05], the *chebfun system* by L. N. Trefethen and coauthors [BT04, PPT10, THP+09], and MATHWORKS' optimization toolbox. These computations are marked by an icon and an *identifier* in the margin of the page and may be reproduced by the reader. The necessary MATLAB files are available upon request by sending a short message to `stefan@guettel.com`.

identifier

2 Functions of Operators

Auch kann $f(U)$ kürzer als das Residuum von $(xE - U)^{-1} f(x)$ in Bezug auf alle Wurzeln der charakteristischen Gleichung von U erklärt werden.

F. G. Frobenius [Fro96]

In this chapter we define operator functions $f(A)$, where f is a complex-valued function and A is a linear operator on a complex Banach space $\mathcal{B}$.

The theory of operator functions includes matrix functions as a special case. However, it took more than 50 years of separate development before both concepts were unified in the early 20th century. Therefore it is scarcely possible to assign precise credit for references to many of the notions we review here. Cayley [Cay58, Cay72] was certainly the first to study square roots of 2×2 and 3×3 matrices. Although published late in 1898, Laguerre [Lag98] had considered infinite power series for constructing the matrix exponential as early as 1867. The definition of a matrix function via polynomial interpolation is due to Sylvester [Syl83] and Buchheim [Buc84, Buc86]. Frobenius [Fro96] stated that if f is analytic then $f(A)$ is the sum of residues of $(\lambda I - A)^{-1} f(\lambda)$ at the eigenvalues of A, attributing an important share of credit to Stickelberger, who used this idea to define powers A^α in his *akademische Antrittsschrift*[1] [Sti81]. Poincaré [Poi99] made explicit use of the Cauchy integral to define $f(A)$ for a matrix. For further details on the interesting history of matrix functions we refer to MacDuffee [Mac46, Ch. IX] and Higham [Hig08, Sec. 1.10].

[1] The *akademische Antrittsschrift* is a paper accompanying Stickelberger's inauguration as professor at the University of Freiburg. In [Tay50] the word *Antrittsschrift* is translated as *dissertation*, and this mistake is often repeated in the literature (e.g., in [DS58, §VII.11]).

The development of the theory of operator functions was mainly stimulated by the need for a spectral theory for bounded operators arising from integral equations, see, e.g., the contributions of Dunford [Dun43] and Taylor [Tay43], and the extension to the unbounded case [Tay50]. For a review of the early historical developments in this direction, and as a general reference for this introductory chapter, we refer to the book by Dunford & Schwartz [DS58, Ch. VII].

In Section 2.1 we consider functions of bounded operators, which are defined by contour integrals. Section 2.2 is devoted to the important special case of algebraic operators. The definition of functions of unbounded operators requires extra care and is discussed in Section 2.3.

2.1 Bounded Operators

Let $\mathcal{B}$ be a complex Banach space. The elements of $\mathcal{B}$ are called *vectors* and the vector norm and associated operator norm are denoted by $\|\cdot\|$. The *domain* of a linear operator A is denoted by $\mathcal{D}(A)$. In this section we assume that A is *bounded on* $\mathcal{B}$, i.e., $\|A\| < \infty$ and $\mathcal{D}(A) = \mathcal{B}$. The space $\mathcal{L}(\mathcal{B})$ of all such operators constitutes a Banach algebra with the usual addition and multiplication operations, zero element O and identity I. The *resolvent set* $\varrho(A)$ is the set of complex numbers ζ for which the *resolvent* $R(\zeta, A) := (\zeta I - A)^{-1}$ exists as an element of $\mathcal{L}(\mathcal{B})$. The *spectrum*[2] $\Lambda(A)$ is the complement of $\varrho(A)$ in $\mathbb{C}$. The following two lemmas are classical in functional analysis, so we just cite them here from [DS58, Ch. VII].

Lemma 2.1 (F. Riesz)**.** *The resolvent set $\varrho(A)$ is open and the resolvent $R(\zeta, A)$ as a function of ζ is analytic in $\varrho(A)$. More precisely, the following power series representation holds:*

$$R(\zeta + \mu, A) = \sum_{j=0}^{\infty} (-\mu)^j R(\zeta, A)^{j+1} \quad \textit{for} \quad |\mu| < \|R(\zeta, A)\|^{-1}.$$

Lemma 2.2 (A. E. Taylor)**.** *The spectrum $\Lambda(A)$ is nonempty and compact.*

We will require some further notions from complex integration (cf. [Hen88, §4.6]). Throughout this thesis, Γ will denote a finite union of nonintersecting piecewise regular Jordan

[2] In the functional analysis literature the spectrum is often denoted by $\sigma(A)$. However, we prefer $\Lambda(A)$, which is also common (e.g., Halmos [Hal72] used it) and more familiar to the linear algebra community when A is a matrix.

curves. To abbreviate this we say that Γ is an *integration contour*. We always assume that integration contours are traversed in the positive sense. A nonempty open subset of $\mathbb{C}$ is called a *region*, and a connected region is a *component*. The set $\mathbb{C} \setminus \Gamma$ consists of two regions, namely the bounded *interior* $\mathrm{int}(\Gamma)$, which is the set of points $z \in \mathbb{C}$ where the *winding number* $n(\Gamma, z) := (2\pi i)^{-1} \int_\Gamma (\zeta - z)^{-1} \, \mathrm{d}\zeta$ equals one, and the unbounded *exterior* $\mathrm{ext}(\Gamma)$ with $n(\Gamma, z) = 0$. Any open inclusion set of a closed set is called a *neighborhood*.

We are now in the position to define operator functions $f(A)$ for functions f analytic in a neighborhood of $\Lambda(A)$.

Definition 2.3. Let Γ be an integration contour such that $\Lambda(A) \subset \mathrm{int}(\Gamma)$, f is analytic in $\mathrm{int}(\Gamma)$ and extends continuously to Γ. Then

$$f(A) := \frac{1}{2\pi i} \int_\Gamma f(\zeta) R(\zeta, A) \, \mathrm{d}\zeta.$$

This formula for $f(A)$ is often called the *Cauchy–Dunford integral*. The integral is defined as the limit of Riemann sums in the operator norm topology. Using Lemma 2.1 and Cauchy's integral theorem [Hen88, Thm. 5.5c] one can easily establish that $f(A)$ is independent of the particular integration contour Γ. Let us collect some important properties of $f(A)$ (cf. [DS58, Ch. VII.3]).

Lemma 2.4 (I. M. Gelfand)**.** *Let f and g be analytic in some neighborhood of $\Lambda(A)$ and let α, β be complex numbers. The following assertions hold.*

(a) $f(A) \in \mathcal{L}(\mathcal{B})$ *and* $g(A) \in \mathcal{L}(\mathcal{B})$,
(b) $\alpha f(A) + \beta g(A) = (\alpha f + \beta g)(A)$,
(c) $f(A) \cdot g(A) = (f \cdot g)(A)$, *hence* $f(A)$ *and* $g(A)$ *commute*,
(d) *if* $f(z) = \sum_{j=0}^{\infty} \alpha_j (z - z_0)^j$ *in a neighborhood of* $\Lambda(A)$,
then $f(A) = \sum_{j=0}^{\infty} \alpha_j (A - z_0 I)^j$,
(e) *if* $B \in \mathcal{L}(\mathcal{B})$ *is invertible, then* $f(BAB^{-1}) = Bf(A)B^{-1}$.

Another useful result is the so-called *spectral mapping theorem*.

Theorem 2.5 (N. Dunford)**.** *If $A \in \mathcal{L}(\mathcal{B})$ and f is analytic in a neighborhood of $\Lambda(A)$, then $\Lambda(f(A)) = f(\Lambda(A))$.*

2.2 Algebraic Operators

An operator $A \in \mathcal{L}(\mathcal{B})$ is *algebraic* if it has a *minimal polynomial* $\psi_A(z) = \prod_{j=1}^{k}(z-\lambda_j)^{m_j}$ (with pairwise distinct λ_j), which is the unique monic polynomial of smallest possible degree satisfying $\psi_A(A) = O$. Square matrices $A \in \mathbb{C}^{N\times N}$ are algebraic operators on $\mathbb{C}^N$. Functions of algebraic operators possess representations alternative to Definition 2.3. By Cauchy's formula for the Hermite interpolation polynomial [Wal69, p. 50], we know that

$$p_{f,A}(z) = \frac{1}{2\pi i}\int_\Gamma \frac{\psi_A(\zeta)-\psi_A(z)}{\zeta - z}\frac{f(\zeta)}{\psi_A(\zeta)}\,\mathrm{d}\zeta$$

is a polynomial that interpolates[3] the function f at the zeros of ψ_A. Using Definition 2.3 we obtain

$$\begin{aligned} p_{f,A}(A) &= \frac{1}{2\pi i}\int_\Gamma [\psi_A(\zeta)I - \psi_A(A)](\zeta I - A)^{-1}\frac{f(\zeta)}{\psi_A(\zeta)}\,\mathrm{d}\zeta \\ &= \frac{1}{2\pi i}\int_\Gamma f(\zeta)R(\zeta,A)\,\mathrm{d}\zeta \\ &= f(A). \end{aligned}$$

Every function of an algebraic operator is therefore representable as a polynomial $p_{f,A}$ depending on f and ψ_A. Let us summarize this finding in the following definition.

Definition 2.6. Let A be algebraic and let the polynomial $p_{f,A}(\lambda)$ satisfy the interpolation conditions

$$p_{f,A}^{(\nu)}(\lambda_j) = f^{(\nu)}(\lambda_j) \qquad \text{for } j = 1,\dots,k \text{ and } \nu = 0,\dots,m_j-1,$$

provided that all $f^{(\nu)}(\lambda_j)$ are defined. Then $f(A) := p_{f,A}(A)$.

Note that f need not be analytic in a neighborhood of $\Lambda(A)$, as is required in Definition 2.3. Instead it is only required that f possesses derivatives up to a finite order. It is clear from Definition 2.6 that two functions f and g satisfy $f(A) = g(A)$ if $\psi_A \mid (p_{f,A} - p_{g,A})$. The converse is also true since $(f-g)(A) = O$ implies $\psi_A \mid p_{f-g,A}$ by the minimality of ψ_A.

An alternative definition of an operator function is based on a *Jordan canonical form* (cf. [Mey00, §7.8]).

[3] In what follows, *interpolation* will always be understood in Hermite's sense.

Definition 2.7. Let $J = U^{-1}AU$ be a Jordan canonical form of $A \in \mathbb{C}^{N\times N}$, where $J = \operatorname{diag}(J_1, \ldots, J_p) \in \mathbb{C}^{N\times N}$ is a matrix with Jordan blocks

$$J_\ell = \begin{bmatrix} \lambda_\ell & 1 & & \\ & \lambda_\ell & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{bmatrix} \in \mathbb{C}^{n_\ell \times n_\ell},$$

and $U \in \mathbb{C}^{N\times N}$ is invertible. Then $f(A) := Uf(J)U^{-1} := U\operatorname{diag}(f(J_1), \ldots, f(J_p))U^{-1}$, where

$$f(J_\ell) := \begin{bmatrix} \frac{f^{(0)}(\lambda_\ell)}{0!} & \frac{f^{(1)}(\lambda_\ell)}{1!} & \cdots & \frac{f^{(n_\ell-1)}(\lambda_\ell)}{(n_\ell-1)!} \\ & \frac{f^{(0)}(\lambda_\ell)}{0!} & \ddots & \vdots \\ & & \ddots & \frac{f^{(1)}(\lambda_\ell)}{1!} \\ & & & \frac{f^{(0)}(\lambda_\ell)}{0!} \end{bmatrix} \quad \text{for } \ell = 1, \ldots, p,$$

provided that all $f^{(\nu)}(\lambda_\ell)$ are defined.

It is easy to show that this definition is independent of the particular Jordan canonical form that is used. For convenience we stated this definition for a matrix A. This is no restriction since every algebraic operator has a matrix representation in an appropriate basis. The equivalence of Definition 2.6 and Definition 2.7 is verified by observing that $f(J_\ell) = p_{f,A}(J_\ell)$ for all $\ell = 1, \ldots, p$. For a detailed review of definitions of matrix functions and their relationships we refer to [Rin55] and [Hig08, Ch. 1].

2.3 Closed Unbounded Operators

We now consider the important class of closed linear operators. An operator A is *closed* if for every sequence $\{\boldsymbol{x}_n\} \subseteq \mathcal{D}(A)$ with $\boldsymbol{x}_n \to \boldsymbol{x}$ and $A\boldsymbol{x}_n \to \boldsymbol{y}$ there holds $\boldsymbol{x} \in \mathcal{D}(A)$ and $A\boldsymbol{x} = \boldsymbol{y}$. Closed operators are more general than bounded operators (e.g., a large class of differential operators is closed but unbounded) but still possess nice enough properties to allow for the development of a functional calculus (under mild additional assumptions).

The definition of the resolvent set and the spectrum of unbounded operators is identical to the corresponding definitions for bounded operators, i.e., $\varrho(A)$ is the set of complex

numbers ζ for which the resolvent $R(\zeta, A) = (\zeta I - A)^{-1}$ exists as a bounded operator with domain $\mathcal{B}$ and $\Lambda(A) = \mathbb{C} \setminus \varrho(A)$. The spectrum $\Lambda(A)$ of a closed unbounded operator is a closed subset of $\mathbb{C}$, which—in contrast to the bounded case—can also be the empty set or even the whole plane $\mathbb{C}$. By the closed graph theorem (cf. [DS58, Thm. II.2.4]) every closed operator A with $\mathcal{D}(A) = \mathcal{B}$ is bounded, hence the domain of a closed unbounded operator is a strict subset of $\mathcal{B}$. This causes technical difficulties when dealing with unbounded operators because one has to ensure that operations with A or $f(A)$ are actually defined. Thus we face the following problem: How to write about rational Krylov methods, which are particularly interesting for unbounded operators, in a way such that generality and rigor are maintained, but the notation remains transparent and close to the one we are used to from linear algebra?

One convenient solution would be to assume that "all operations involving A are defined." However, we found such an assumption too sloppy. A (hopefully) satisfying compromise for this thesis is to consider bounded linear operators only, referring to the fact that if there exists a number $\xi \in \varrho(A)$, the unbounded operator A can be transformed into a bounded operator $(A - \xi I)^{-1} \in \mathcal{L}(\mathcal{B})$. The definition

$$f(A) := \hat{f}((A - \xi I)^{-1}), \quad \text{with } \hat{f}(z) := f(z^{-1} + \xi), \tag{2.1}$$

is thus reduced to a function of a bounded operator. This way of defining $f(A)$ for closed unbounded operators with nonempty resolvent set was proposed by Taylor [Tay50] and is standard now (cf. [DS58, §VII.9]). The assumption that A be closed is essential to build a functional calculus on (2.1). For example, if A is closed and has a nonempty resolvent set, then $p(A)$ is a closed operator for every polynomial p [DS58, Thm. VII.9.7].

Note that the transformation (2.1) is the same as the one in the shift-and-invert Lanczos method considered in the introduction (page 3ff.). There we actually transformed a finite-difference matrix A with spectral interval contained in $(-\infty, 0)$ into a matrix $(A - \xi I)^{-1}$ with bounded spectral interval contained in $(-1/\xi, 0)$. Since the spectral interval of A becomes arbitrarily large as the discretization mesh becomes finer, it is adequate to consider A as an unbounded operator.

3 Subspace Approximation for $f(A)\boldsymbol{b}$

Oscillatory movement is gaining more and more importance in technical problems, so that in Germany there is a study program called Schwingungs-Ingenieure.

A. N. Krylov [Kry31]

Let A be a bounded linear operator on a complex Hilbert space $\mathcal{H}$ and let f be a complex-valued function such that $f(A)$ is defined. Let a vector $\boldsymbol{b} \in \mathcal{H}$ be given. Our aim is to obtain an approximation for $f(A)\boldsymbol{b}$ from a subspace $\mathcal{V}_m \subseteq \mathcal{H}$ of small dimension m, while avoiding the explicit "evaluation" of $f(A)$, which is usually unfeasible or even impossible.

In Section 3.1 we consider the so-called Rayleigh method, which is a general method for obtaining approximations for $f(A)\boldsymbol{b}$. It turns out that this method applied to the solution of linear operator equations is closely related to the Galerkin method. In Section 3.2 we explore the fact that Rayleigh approximations can be interpreted as linear combinations of so-called Ritz vectors of A. Finally, in Section 3.3 we consider the important special case where the search space $\mathcal{V}_m$ is a polynomial Krylov space and collect results about the associated Rayleigh–Ritz approximations.

3.1 The Rayleigh Method

Let $\mathcal{H}$ be endowed with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\boldsymbol{h}\| = \langle \boldsymbol{h}, \boldsymbol{h} \rangle^{1/2}$.

We consider a *quasi-matrix*[1] $V_m := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$ whose columns $\boldsymbol{v}_j$ form a basis of an m-dimensional subspace $\mathcal{V}_m \subseteq \mathcal{H}$. For an arbitrary vector $\boldsymbol{c} = [c_1, \ldots, c_m]^T \in \mathbb{C}^m$ we define

$$V_m \boldsymbol{c} := \boldsymbol{v}_1 c_1 + \cdots + \boldsymbol{v}_m c_m,$$

so that $V_m : \mathbb{C}^m \to \mathcal{H}$ is a linear operator. The *coordinate space* $\mathbb{C}^m$ is always endowed with the Euclidian inner product. Using the triangle inequality it is readily verified that the operator norm of V_m satisfies $\|V_m\| \leq \sum_{j=1}^m \|\boldsymbol{v}_j\|$, i.e., V_m is a bounded operator. Since the *range* $\mathcal{R}(V_m) = \mathcal{V}_m$ is finite-dimensional and therefore closed, there exists a unique *Moore–Penrose inverse* $V_m^\dagger : \mathcal{H} \to \mathbb{C}^m$, which is uniquely defined as the solution X of the four equations [BG03, Ch. 9, Thm. 3]

$$V_m X V_m = V_m, \quad X V_m X = X, \quad (X V_m)^* = X V_m, \quad (V_m X)^* = V_m X.$$

By the first of these equations $(V_m V_m^\dagger)^2 = V_m V_m^\dagger$ and by the last $(V_m V_m^\dagger)^* = V_m V_m^\dagger$, so that $V_m V_m^\dagger$ is the orthogonal projection operator onto $\mathcal{V}_m$, and therefore

$$\langle V_m V_m^\dagger \boldsymbol{h}, \boldsymbol{v} \rangle = \langle \boldsymbol{h}, \boldsymbol{v} \rangle \quad \text{for all } \boldsymbol{h} \in \mathcal{H},\ \boldsymbol{v} \in \mathcal{V}_m. \tag{3.1}$$

Since V_m has full column rank, we have two more useful properties of the Moore–Penrose inverse at hand:

$$(V_m S)^\dagger = S^{-1} V_m^\dagger \quad \text{for every invertible } S \in \mathbb{C}^{m \times m}, \tag{3.2}$$

$$V_m^\dagger V_m = I_m, \quad \text{where } I_m \text{ denotes the } m \times m \text{ identity matrix.} \tag{3.3}$$

Equation (3.2) follows from the fact that $X = (V_m S)^\dagger$ solves $(V_m S) X (V_m S) = (V_m S)$, or equivalently, $V_m (S X) V_m = V_m$. Hence $V_m^\dagger = SX = S(V_m S)^\dagger$. Equation (3.3) holds because $V_m V_m^\dagger \boldsymbol{v}_j = \boldsymbol{v}_j$ and hence $V_m^\dagger \boldsymbol{v}_j = \boldsymbol{e}_j$, the jth column of I_m.

[1] This term was coined by Stewart [Ste98, Ch. 5]. However, the same idea with different terminology was used before in [Boo91] and [TB97, p. 52].

Let A be a bounded linear operator on $\mathcal{H}$. The following two definitions are fundamental.

Definition 3.1. For a given quasi-matrix $V_m = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$ of full column rank, the *Rayleigh quotient for* (A, V_m) is defined as

$$A_m := V_m^\dagger A V_m = [V_m^\dagger(A\boldsymbol{v}_1), \ldots, V_m^\dagger(A\boldsymbol{v}_m)] \in \mathbb{C}^{m\times m}.$$

The name *Rayleigh quotient* is justified since A_m is the generalization of the so-called *matrix Rayleigh quotient*, which results when the columns of V_m are orthonormal vectors (cf. [Par98, §11.3]).

Definition 3.2. Let V_m be a basis[2] of $\mathcal{V}_m$ and denote by A_m the Rayleigh quotient for (A, V_m). Provided that $f(A_m)$ exists, the *Rayleigh approximation for* $f(A)\boldsymbol{b}$ *from* $\mathcal{V}_m$ is defined as

$$\boldsymbol{f}_m := V_m f(A_m) V_m^\dagger \boldsymbol{b}.$$

To justify this definition we need to verify that $\boldsymbol{f}_m$ is independent of the choice of the basis V_m.

Lemma 3.3. *Let V_m and W_m be bases of $\mathcal{V}_m$. Then*

(a) *the Rayleigh quotients for (A, V_m) and (A, W_m) are similar matrices,*

(b) *the Rayleigh approximation is independent of the particular choice of the basis and depends only on the search space $\mathcal{V}_m$.*

Proof. There exists an invertible matrix $S \in \mathbb{C}^{m\times m}$ such that $W_m = V_m S$. Therefore

(a) $A'_m := W_m^\dagger A W_m = S^{-1} V_m^\dagger A V_m S = S^{-1} A_m S$ using (3.2), and

(b) $W_m f(A'_m) W_m^\dagger \boldsymbol{b} = V_m S f(S^{-1} A_m S) S^{-1} V_m^\dagger \boldsymbol{b} = V_m f(A_m) V_m^\dagger \boldsymbol{b}$, where we have used the property $f(S^{-1} A_m S) = S^{-1} f(A_m) S$ (which holds for every matrix function). □

For a given search space $\mathcal{V}_m$ a method for obtaining a Rayleigh approximation is referred to as a *Rayleigh method*.

[2] This is a short-hand for "Let the columns of $V_m = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$ be a basis of ..."

Remark 3.4. The term $f(A_m)$ in the definition of the Rayleigh approximation is a function of a (relatively) small matrix, whereas evaluating $f(A)$ is usually unfeasible or even impossible (for example if A is defined only by its action on vectors). For a comprehensive treatment and an extensive bibliography of the $f(A_m)$ problem we refer to the recent monograph by Higham [Hig08] and the review by Frommer & Simoncini [FS08a].

Remark 3.5. Assume that A is invertible. The Rayleigh approximation for $f(z) = z^{-1}$ coincides with the approximation obtained by the *(Bubnov–)Galerkin method* applied to the operator equation $A\boldsymbol{x} = \boldsymbol{b}$ (cf. [Kre99, §13.3]). To show this we verify that the Rayleigh approximation $\boldsymbol{x}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b}$ satisfies the Galerkin condition. In the context of operator equations this means that the residual is orthogonal to the *test space* $\mathcal{V}_m$, i.e.,

$$\langle \boldsymbol{b} - A\boldsymbol{x}_m, \boldsymbol{v} \rangle = 0 \quad \text{for all } \boldsymbol{v} \in \mathcal{V}_m. \tag{3.4}$$

Inserting the Rayleigh approximation yields

$$\begin{aligned} \langle \boldsymbol{b} - A\boldsymbol{x}_m, \boldsymbol{v} \rangle &= \langle \boldsymbol{b} - AV_m A_m^{-1} V_m^\dagger \boldsymbol{b}, \boldsymbol{v} \rangle \\ &= \langle V_m V_m^\dagger \boldsymbol{b} - V_m (V_m^\dagger A V_m) A_m^{-1} V_m^\dagger \boldsymbol{b}, \boldsymbol{v} \rangle \\ &= 0 \quad \text{for all } \boldsymbol{v} \in \mathcal{V}_m, \end{aligned}$$

where we have used (3.1) for the second equality. Conversely, there is no other vector $\boldsymbol{x}'_m \in \mathcal{V}_m$ that satisfies the Galerkin condition (3.4): by linearity of the inner product such a vector satisfies $\langle A(\boldsymbol{x}_m - \boldsymbol{x}'_m), \boldsymbol{v} \rangle = 0$, and by (3.1), $\langle V_m V_m^\dagger A(\boldsymbol{x}_m - \boldsymbol{x}'_m), \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{v} \in \mathcal{V}_m$. Setting $\boldsymbol{v} = V_m V_m^\dagger A_m (\boldsymbol{x}_m - \boldsymbol{x}'_m) \in \mathcal{V}_m$ and using the fact that A is injective, we obtain $\boldsymbol{x}_m = \boldsymbol{x}'_m$. Therefore the Galerkin condition (3.4) completely characterizes the Rayleigh approximation from Definition 3.2 for the function $f(z) = z^{-1}$.

For other functions f a residual equation is usually not available and the Galerkin method is not applicable. However, the Rayleigh approximation is defined whenever $f(A_m)$ is defined.

3.2 Ritz Pairs

Definition 3.6. Let $(\theta_j, \boldsymbol{x}_j)$ be an eigenpair of the Rayleigh quotient $A_m = V_m^\dagger A V_m$, i.e., $A_m \boldsymbol{x}_j = \boldsymbol{x}_j \theta_j$. Then $(\theta_j, \boldsymbol{y}_j := V_m \boldsymbol{x}_j)$ is a *Ritz pair for* $(A, \mathcal{V}_m)$. The number $\theta_j \in \mathbb{C}$ is called a *Ritz value* and $\boldsymbol{y}_j \in \mathcal{V}_m$ is the associated *Ritz vector*.

By Lemma 3.3 the Ritz values are indeed independent of the particular basis V_m of $\mathcal{V}_m$. Since the Ritz values are eigenvalues of a Rayleigh quotient of A, they are all contained in the *numerical range*

$$\mathbb{W}(A) := \left\{ \frac{\langle A\boldsymbol{h}, \boldsymbol{h}\rangle}{\langle \boldsymbol{h}, \boldsymbol{h}\rangle} \,:\, \boldsymbol{h} \in \mathcal{H} \setminus \{\boldsymbol{0}\} \right\},$$

which, by the *Toeplitz–Hausdorff theorem* (cf. [Hal82, Ch. 22]), is a bounded convex subset of $\mathbb{C}$ whose closure contains the spectrum $\Lambda(A)$.

It is easily verified that also the Ritz vectors are independent of the basis V_m, hence we can assume that V_m is an orthonormal basis. In this case the Rayleigh quotient and the coordinate representation of $\boldsymbol{b}$ simplify to

$$A_m = V_m^* A V_m = [\langle A\boldsymbol{v}_\ell, \boldsymbol{v}_k\rangle]_{1\le k,\ell\le m} \quad \text{and} \quad V_m^\dagger \boldsymbol{b} = V_m^* \boldsymbol{b} = [\langle \boldsymbol{b}, \boldsymbol{v}_k\rangle]_{1\le k\le m}. \tag{3.5}$$

In Remark 3.5 we have discussed the relationship between the Rayleigh method and the Galerkin method for solving operator equations. In fact, a connection to operator eigenproblems can also be given. To this end, let us consider a self-adjoint operator $A = A^*$. Then the matrix A_m in (3.5) is Hermitian, and hence has m orthonormal eigenvectors $\boldsymbol{x}_j$ associated with real Ritz values θ_j, i.e.,

$$A_m \boldsymbol{x}_j = \boldsymbol{x}_j \theta_j \quad \text{for } j = 1, \ldots, m. \tag{3.6}$$

The Ritz vectors $\boldsymbol{y}_j = V_m \boldsymbol{x}_j$ form an orthonormal basis of $\mathcal{V}_m$. Writing $\boldsymbol{x}_j$ component-wise $\boldsymbol{x}_j = [x_{1,j}, \ldots, x_{m,j}]^T$, the kth row of (3.6) reads as

$$\sum_{\ell=1}^{m} \langle A\boldsymbol{v}_\ell, \boldsymbol{v}_k\rangle x_{\ell,j} - x_{k,j}\theta_j = 0,$$

or equivalently, $\langle (A-\theta_j I)\boldsymbol{y}_j, \boldsymbol{v}_k\rangle = 0$. Varying $k = 1,\dots,m$, we arrive at the variational formulation

$$\langle (A-\theta_j I)\boldsymbol{y}_j, \boldsymbol{v}\rangle = 0 \quad \text{for all } \boldsymbol{v} \in \mathcal{V}_m.$$

This means that the Ritz vectors $\boldsymbol{y}_j \in \mathcal{V}_m$ may be regarded as "eigenvector approximations" and the θ_j are corresponding approximate "eigenvalues" (A itself may not have any eigenvalues). Moreover, the Rayleigh approximation obviously satisfies

$$\boldsymbol{f}_m = V_m f(A_m) V_m^* \boldsymbol{b} = \sum_{j=1}^{m} f(\theta_j)\langle \boldsymbol{b}, \boldsymbol{y}_j\rangle \boldsymbol{y}_j,$$

meaning that $f(A)\boldsymbol{b}$ is approximated by a linear combination of approximate eigenvectors $\boldsymbol{y}_j$ of A scaled by $f(\theta_j)$ ($j = 1,\dots,m$).

3.3 Polynomial Krylov Spaces

In 1931 A. N. Krylov published the paper [Kry31] in which he considered the problem of computing eigenvalues of a square matrix $A \in \mathbb{C}^{N\times N}$. Starting with some vector $\boldsymbol{b} \in \mathbb{C}^N$, Krylov used the sequence $\{\boldsymbol{b}, A\boldsymbol{b}, \dots, A^{N-1}\boldsymbol{b}\}$ in a clever way to find the coefficients of the characteristic polynomial $\chi_A(z)$ of A with significantly fewer arithmetic operations than the direct expansion of the determinant $\det(zI - A)$ would require.[3] Today this attempt is viewed as an "unfortunate goal" [Par98, Ch. 12], as it is known to be a highly unstable approach for computing eigenvalues. However, back in 1931 people were dealing with, say, 6×6 matrices where ill-conditioning does not play such a big role [Bot02]. Later, Krylov's name became attached to the spaces we now call *polynomial Krylov spaces*: for a bounded linear operator A and a vector $\boldsymbol{b} \in \mathcal{H}$ the polynomial Krylov space of order m associated with $(A, \boldsymbol{b})$ is

$$\mathcal{K}_m(A, \boldsymbol{b}) = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, \dots, A^{m-1}\boldsymbol{b}\}.$$

If there is no room for ambiguity, we will often write $\mathcal{K}_m$ instead of $\mathcal{K}_m(A, \boldsymbol{b})$. With increasing order m, polynomial Krylov spaces are nested subspaces of $\mathcal{H}$.

Polynomial approximation methods for the $f(A)\boldsymbol{b}$ problem started to gain interest in the mid 1980's. Until that time the most common approach to solve such problems was by

[3] see [FF76, §42] for a brief summary of Krylov's method. The computation of χ_A is only possible if A is nonderogatory and $\boldsymbol{b}$ is cyclic for A, otherwise Krylov's method computes a divisor of χ_A.

diagonalization of A (with exception of the function $f(z) = z^{-1}$ for which efficient polynomial Krylov methods had already been known for about 30 years, see [SV00, §6]). Nauts & Wyatt [NW83, NW84], interested in the scalar problem $\langle \exp(A)\boldsymbol{b}, \boldsymbol{a}\rangle$ for Hamiltonian operators A arising in Chemical Physics and vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, realized that full diagonalization is not necessary if one is able to compute a few eigenvalues of A by the Lanczos method. Building on this work, Park & Light [PL86] advocated the efficiency of polynomial Krylov methods for $\exp(A)\boldsymbol{b}$, making use of Rayleigh approximations. Since then, polynomial Krylov methods were further applied and analyzed for the matrix exponential [Nou87, FTDR89], and more generally for other functions $f(z)$ [Vor87, DK89, Kni91]. The theoretical understanding of these methods was greatly enhanced by Ericsson [Eri90] and Saad [Saa92a], who independently discovered the connection between Rayleigh approximations and polynomial interpolation.

In this section the search space for the Rayleigh method is a polynomial Krylov space, i.e., $\mathcal{V}_m = \mathcal{K}_m$. Let us collect some well-known properties of the associated Rayleigh quotients in the following lemma (cf. [Saa92b, Ch. VI]). By $\mathcal{P}_m$ we denote the space of polynomials of degree $\leq m$, and $\mathcal{P}_m^\infty$ denotes the set of monic polynomials of degree $= m$.

Lemma 3.7. *Let V_m be a basis of $\mathcal{K}_m(A, \boldsymbol{b})$, $A_m = V_m^\dagger A V_m$, and let χ_m denote the characteristic polynomial of A_m. Then the following statements hold.*

(a) *A_m is nonderogatory.*

(b) *$\chi_m(A)\boldsymbol{b} \perp \mathcal{K}_m(A, \boldsymbol{b})$.*

(c) *χ_m minimizes $\|s_m(A)\boldsymbol{b}\|$ among all $s_m \in \mathcal{P}_m^\infty$.*

Proof.

(a) Assume first that $V_m = [\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}]$. Then $A_m = V_m^\dagger A V_m$ is obviously a companion matrix, i.e.,

$$A_m = \begin{bmatrix} 0 & & & -\alpha_0 \\ 1 & 0 & & -\alpha_1 \\ & \ddots & \ddots & \vdots \\ & & 1 & -\alpha_{m-1} \end{bmatrix},$$

with characteristic polynomial $\chi_m(z) = z^m + \sum_{j=0}^{m-1} \alpha_j z^j$. Since a companion matrix

is nonderogatory (cf. [Mey00, p. 648]) and for any choice of V_m the matrix A_m is similar to this companion matrix (by Lemma 3.3), A_m is nonderogatory.

(b) Assume again that $V_m = [\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}]$. Then $V_m^\dagger A^m \boldsymbol{b} = [-\alpha_0, -\alpha_1, \ldots, -\alpha_{m-1}]^T$, which is the last column vector of A_m. Since $V_m V_m^\dagger A^j \boldsymbol{b} = A^j \boldsymbol{b}$ for all $j \leq m-1$, we have

$$V_m V_m^\dagger \chi_m(A)\boldsymbol{b} = V_m V_m^\dagger A^m \boldsymbol{b} + \sum_{j=0}^{m-1} \alpha_j A^j \boldsymbol{b} = \boldsymbol{0},$$

as desired.

(c) Write $s_m(z) = z^m - p_{m-1}(z)$ for some $p_{m-1} \in \mathcal{P}_{m-1}$. Then $p_{m-1}(A)\boldsymbol{b} \in \mathcal{K}_m$ and the condition

$$\langle A^m \boldsymbol{b} - p_{m-1}(A)\boldsymbol{b}, \boldsymbol{v} \rangle \quad \text{for all} \quad \boldsymbol{v} \in \mathcal{K}_m$$

characterizes the unique minimizer of $\|s_m(A)\boldsymbol{b}\|$. By (b) this condition is satisfied for $s_m = \chi_m$. □

We now turn to Rayleigh approximations from polynomial Krylov spaces, which deserve a special name.

Definition 3.8. The Rayleigh approximation

$$\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b}$$

is called a *Rayleigh–Ritz approximation for* $f(A)\boldsymbol{b}$ if V_m is a basis of $\mathcal{K}_m(A, \boldsymbol{b})$.

We will see that the Ritz values $\Lambda(A_m)$ have a very important meaning in connection with Rayleigh–Ritz approximations—hence the name. In fact, the Ritz values turn out to be interpolation nodes for a polynomial underlying $\boldsymbol{f}_m$. In [HH05] the authors used the name *Ritz approximation* for $\boldsymbol{f}_m$, but we intend to remind the reader that Rayleigh–Ritz approximations are Rayleigh approximations from a special search space.

In Remark 3.5 we showed that the Rayleigh method applied for the function $f(z) = z^{-1}$ coincides with the Galerkin method for the solution of an operator equation, and consequently the same is true for the Rayleigh–Ritz method. A Galerkin method with search space $\mathcal{V}_m = \mathcal{K}_m$ is also known as *method of moments* (cf. [Vor65]).

The following lemma is a straightforward generalization of [Eri90, Thm. 5.1] and [Saa92a, Lem. 3.1]; we have dropped the assumptions that V_m is orthonormal and that A is a matrix.

Lemma 3.9 (Ericsson, Saad)**.** *Let V_m be a basis of $\mathcal{K}_m$, $A_m = V_m^\dagger A V_m$, and let $P_m = V_m V_m^\dagger$ be the orthogonal projector onto $\mathcal{K}_m$. Then for every $p_m \in \mathcal{P}_m$ there holds*

$$P_m p_m(A)\boldsymbol{b} = V_m p_m(A_m) V_m^\dagger \boldsymbol{b}.$$

In particular, for every $p_{m-1} \in \mathcal{P}_{m-1}$ there holds

$$p_{m-1}(A)\boldsymbol{b} = V_m p_{m-1}(A_m) V_m^\dagger \boldsymbol{b},$$

i.e., the Rayleigh–Ritz approximation for $p_{m-1}(A)\boldsymbol{b}$ is exact.

Proof. The proof is by induction on the monomials A^j, i.e., we will show that $P_m A^j \boldsymbol{b} = V_m A_m^j V_m^\dagger \boldsymbol{b}$ for all $j \leq m$. This assertion is obviously true for $j = 0$. Assume that it is true for some $j \leq m-1$. Since $A^j\boldsymbol{b}$ belongs to $\mathcal{K}_m$ we have

$$P_m A^{j+1}\boldsymbol{b} = P_m A A^j \boldsymbol{b} = P_m A P_m A^j \boldsymbol{b}.$$

By definition of A_m we have $P_m A P_m = V_m A_m V_m^\dagger$ and, using the induction hypothesis,

$$P_m A^{j+1}\boldsymbol{b} = (V_m A_m V_m^\dagger) V_m A_m^j V_m^\dagger \boldsymbol{b} = V_m A_m^{j+1} V_m^\dagger \boldsymbol{b}. \qquad \square$$

Of great importance for the analysis of Rayleigh–Ritz approximations is the following theorem, which is a generalization of [Eri90, Thm. 4.3] and [Saa92a, Thm. 3.3]. It also justifies the name *Rayleigh–Ritz approximation*.

Theorem 3.10 (Ericsson, Saad)**.** *Let V_m be a basis of $\mathcal{K}_m$ and $A_m = V_m^\dagger A V_m$. Let f be a function such that $f(A_m)$ is defined. Then*

$$\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b} = p_{m-1}(A)\boldsymbol{b},$$

where $p_{m-1} \in \mathcal{P}_{m-1}$ interpolates f at the Ritz values $\Lambda(A_m)$.

Proof. By the definition of a matrix function, $f(A_m) = p_{m-1}(A_m)$. Hence,

$$V_m f(A_m) V_m^\dagger \boldsymbol{b} = V_m p_{m-1}(A_m) V_m^\dagger \boldsymbol{b} = p_{m-1}(A)\boldsymbol{b},$$

where we have used Lemma 3.9 for the last equality. □

Let us further investigate in what case the Rayleigh–Ritz approximation is exact, i.e., $\boldsymbol{f}_m = f(A)\boldsymbol{b}$. By M we denote the smallest integer such that $\mathcal{K}_{M-1} \subset \mathcal{K}_M = \mathcal{K}_{M+1}$, which means that $\mathcal{K}_M$ is A-invariant (M is called the *invariance index*). If there exists no such integer we set $M = \infty$. If $M < \infty$ there exists a unique polynomial $\psi_{A,\boldsymbol{b}} \in \mathcal{P}_M^\infty$ such that $\psi_{A,\boldsymbol{b}}(A)\boldsymbol{b} = \boldsymbol{0}$, and there exists no such polynomial of smaller degree. This polynomial $\psi_{A,\boldsymbol{b}}$ is called the *minimal polynomial of* $\boldsymbol{b}$ *with respect to* A (cf. [Gan59, Ch. VII]). Let V_M be an arbitrary basis of $\mathcal{K}_M$ and let χ_M denote the characteristic polynomial of the Rayleigh quotient $A_M = V_M^\dagger A V_M$. With the help of Lemma 3.9 we derive

$$\mathcal{K}_M \ni \chi_M(A)\boldsymbol{b} = P_M \chi_M(A)\boldsymbol{b} = V_M \chi_M(A_M) V_M^\dagger \boldsymbol{b} = \boldsymbol{0},$$

which immediately proves

$$\psi_{A,\boldsymbol{b}} = \chi_M = \psi_M,$$

where ψ_M denotes the minimal polynomial of A_M (which coincides with χ_M because A_M is nonderogatory by Lemma 3.7).

By $A^{(M)}$ we denote the *section of* A *onto* $\mathcal{K}_M$, that is $A^{(M)} := V_M A_M V_M^\dagger$. It is easily verified that the Rayleigh quotients A_m for (A, V_m) and $A_m^{(M)}$ for $(A^{(M)}, V_m)$ coincide for all orders $m \leq M$: since $V_M V_M^\dagger$ is a projection onto $\mathcal{K}_M$ there holds

$$A_m = V_m^\dagger A V_m = V_m^\dagger (V_M V_M^\dagger) A (V_M V_M^\dagger) V_m = V_m^\dagger A^{(M)} V_m = A_m^{(M)}. \tag{3.7}$$

This implies that also the Rayleigh–Ritz approximations for $f(A)\boldsymbol{b}$ and $f(A^{(M)})\boldsymbol{b}$ from $\mathcal{K}_m$ coincide (provided that both exist), i.e.,

$$\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b} = V_m f(A_m^{(M)}) V_m^\dagger \boldsymbol{b} \quad \text{for all } m \leq M. \tag{3.8}$$

The following lemma is now easily proved.

Lemma 3.11. *If $M < \infty$ and $f(A)\boldsymbol{b} = f(A^{(M)})\boldsymbol{b}$, then*

$$\boldsymbol{f}_M = V_M f(A_M) V_M^\dagger \boldsymbol{b} = f(A)\boldsymbol{b}.$$

In particular, if A is algebraic then $\boldsymbol{f}_M = f(A)\boldsymbol{b}$.

Proof. By (3.8) we know that $\boldsymbol{f}_M = V_M f(A_M^{(M)}) V_M^\dagger \boldsymbol{b}$, and hence it suffices to show that $V_M f(A_M^{(M)}) V_M^\dagger \boldsymbol{b} = f(A^{(M)})\boldsymbol{b}$, or equivalently, $V_M f(A_M) V_M^\dagger \boldsymbol{b} = f(A^{(M)})\boldsymbol{b}$ (because $A_M = A_M^{(M)}$ by (3.7)). Note that $A^{(M)}$ is an algebraic operator since ψ_M, the minimal polynomial of A_M, satisfies

$$\psi_M(A^{(M)}) = \psi_M(V_M A_M V_M^\dagger) = V_M \psi_M(A_M) V_M^\dagger = O.$$

Hence there exists a polynomial $p_{M-1} \in \mathcal{P}_{M-1}$ that interpolates f at the zeros of ψ_M such that $f(A^{(M)}) = p_{M-1}(A^{(M)})$, and this polynomial also satisfies $f(A_M) = p_{M-1}(A_M)$. Therefore

$$f(A^{(M)})\boldsymbol{b} = p_{M-1}(A^{(M)})\boldsymbol{b} = V_M p_{M-1}(A_M) V_M^\dagger \boldsymbol{b} = V_M f(A_M) V_M^\dagger \boldsymbol{b},$$

where we have used the exactness of Rayleigh–Ritz approximation, Lemma 3.9, for the second equality.

If A is algebraic, then by definition of a polynomial Krylov space it is clear that M is finite[4], in particular, there holds $M \leq \deg(\psi_A)$. That $f(A)\boldsymbol{b} = f(A^{(M)})\boldsymbol{b}$ follows from the fact that the minimal polynomials of $\boldsymbol{b}$ with respect to A and $A^{(M)}$ coincide. □

Remark 3.12. In general, the Rayleigh–Ritz approximation $\boldsymbol{f}_m$ differs from $f(A)\boldsymbol{b}$ until the invariance index M is reached, even if $f(A)\boldsymbol{b} \in \mathcal{K}_m$ for $m < M$. In Figure 3.1 this fact is illustrated with the help of a simple example.

[4] In a sense, a theorem by I. Kaplansky states that the converse is also true: if there exists an integer $M < \infty$ such that $\mathcal{K}_M(A, \boldsymbol{h}) = \mathcal{K}_{M+1}(A, \boldsymbol{h})$ *for every* $\boldsymbol{h} \in \mathcal{H}$ then A is algebraic [Nev93, p. 38].

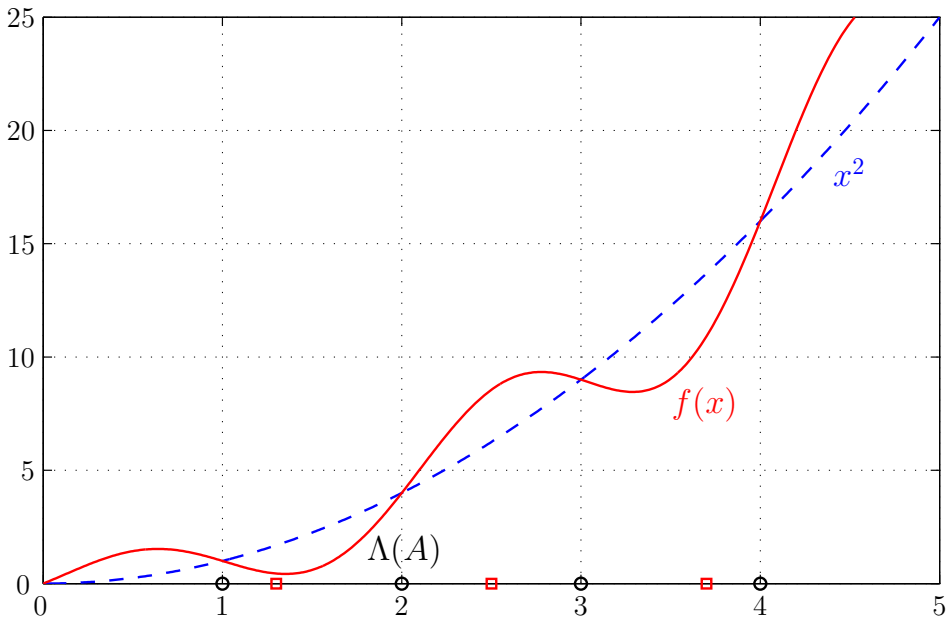


Figure 3.1: The black circles indicate the eigenvalues of $A = \mathrm{diag}(1, 2, 3, 4)$. By chance, $f(A)$ coincides with A^2, so that $f(A)\boldsymbol{b} \in \mathcal{K}_3(A, \boldsymbol{b}) = \mathrm{span}\{\boldsymbol{b}, A\boldsymbol{b}, A^2\boldsymbol{b}\}$. However, the Rayleigh–Ritz approximation of order 3 corresponds to an interpolating polynomial for the function f at the Ritz values (red squares) and thus $\boldsymbol{f}_3$ does not coincide with $f(A)\boldsymbol{b}$.

4 Rational Krylov Spaces

It is thus well possible that the series of largest roots in which we are primarily interested is practically established with sufficient accuracy after a few iterations.

C. Lanczos [Lan50]

Some of the most successful iterative algorithms for eigenproblems with a large sparse or structured matrix A originate from ideas of Lanczos [Lan50] and Arnoldi [Arn51], and are now known as Hermitian Lanczos, non-Hermitian Lanczos, and Arnoldi method, respectively (cf. [Saa92b, Par98]). These methods are polynomial Krylov methods[1] because the computed eigenvector approximations are elements of the polynomial Krylov space $\mathcal{K}_m(A, \boldsymbol{b})$. Lanczos already realized that his method usually tends to approximate *extremal* eigenvalues of symmetric matrices, i.e., those eigenvalues close to the endpoints of A's spectral interval, after a few iterations. Ericsson & Ruhe [ER80] made use of this observation by running the Lanczos method for the spectrally transformed matrix $(A - \xi I)^{-1}$, thus obtaining good approximate eigenvalues close to the shift $\xi \in \mathbb{C} \setminus \Lambda(A)$. The authors also provided an implementation of this method called STLM (for Spectral Transformation Lanczos Method [ER82]). The search space in the mth iteration of this method is the linear space of rational functions in A times $\boldsymbol{b}$ having a pole of order $m-1$ in ξ, i.e.,

$$\{r_m(A)\boldsymbol{b} \,:\, r_m(z) = p_{m-1}(z)/(z-\xi)^{m-1},\, p_{m-1} \in \mathcal{P}_{m-1}\}.$$

[1] A footnote in [Lan50] indicates that this relationship with Krylov's work was pointed out to Lanczos by A. M. Ostrowski.

Ruhe [Ruh84] developed this idea further by allowing the $m-1$ poles of the rational functions to be arbitrary (not lying in $\Lambda(A)$, of course). The resulting *rational Krylov space* is the linear space of the form

$$\{r_m(A)\boldsymbol{b} \,:\, r_m(z) = p_{m-1}(z)/q_{m-1}(z),\, p_{m-1} \in \mathcal{P}_{m-1}\},$$

where $q_{m-1} \in \mathcal{P}_{m-1}$ is a prescribed polynomial. Such spaces are the subject of this chapter and we will see that they are powerful search spaces for approximating $f(A)\boldsymbol{b}$.

In Section 4.1 we show some basic properties of rational Krylov spaces, where we use a definition based on polynomial Krylov spaces. This allows us to easily carry over all results about polynomial Krylov spaces (cf. Section 3.3) to the rational case. Although this approach inevitably introduces some redundancies into the development of the theory, it most clearly reveals the relationships between polynomial and rational Krylov spaces. In Section 4.2 we show that Rayleigh approximations for $f(A)\boldsymbol{b}$ from rational Krylov spaces are closely related to rational interpolation at rational Ritz values and possess a near-optimality property.

4.1 Definition and Basic Properties

Let A be a bounded linear operator. The following definition and notation will be used.

Definition 4.1. Let the polynomial $q_{m-1} \in \mathcal{P}_{m-1}$ have no zeros in the spectrum $\Lambda(A)$. The space

$$\mathcal{Q}_m := \mathcal{Q}_m(A, \boldsymbol{b}) := q_{m-1}(A)^{-1}\mathcal{K}_m(A, \boldsymbol{b})$$

is called the *rational Krylov space of order m associated with* $(A, \boldsymbol{b}, q_{m-1})$.

The name "$\mathcal{Q}_m$" is intended to remind the reader that there is *always* associated with it a polynomial q_{m-1}, even if this polynomial does not occur explicitly in the notation. In what follows we denote by $\mathcal{P}_{m-1}/q_{m-1}$ the set of rational functions $\{p_{m-1}/q_{m-1} \,:\, p_{m-1} \in \mathcal{P}_{m-1}\}$ and we use a similar notation with $\mathcal{P}_m/q_{m-1}$ and $\mathcal{P}_m^\infty/q_{m-1}$. Let us collect some basic properties of the rational Krylov space $\mathcal{Q}_m$.

Lemma 4.2. *There holds*

(a) $\mathcal{Q}_m = \mathcal{K}_m(A, q_{m-1}(A)^{-1}\boldsymbol{b})$,

(b) $\boldsymbol{b} \in \mathcal{Q}_m$,

(c) $\dim(\mathcal{Q}_m) = \min\{m, M\}$, *where* M *is the invariance index of* $\mathcal{K}_m(A, \boldsymbol{b})$,

(d) $\mathcal{Q}_m \cong \mathcal{P}_{m-1}/q_{m-1}$ *for all* $m \leq M$, *i.e., there exists an isomorphism.*

Proof.

(a) This is a consequence of the fact that operator functions commute. More precisely, $q_{m-1}(A)^{-1}A^j\boldsymbol{b} = A^j q_{m-1}(A)^{-1}\boldsymbol{b}$ for all $j \geq 0$.

(b) We have $\boldsymbol{b} \in q_{m-1}(A)^{-1}\mathcal{K}_m$ if and only if $q_{m-1}(A)\boldsymbol{b} \in \mathcal{K}_m$, the latter being true by definition of $\mathcal{K}_m$.

(c) By definition, $q_{m-1}(A)^{-1}$ is an invertible operator and hence does not reduce the dimension of the space it is applied to. Thus $\dim(q_{m-1}(A)^{-1}\mathcal{K}_m) = \dim(\mathcal{K}_m)$.

(d) This is a consequence of the fact $\mathcal{K}_m \cong \mathcal{P}_{m-1}$. □

We emphasize that assertion (a) of Lemma 4.2 is as important as it is trivial since it links rational and polynomial Krylov spaces via the modified starting vector $q_{m-1}(A)^{-1}\boldsymbol{b}$. We will make extensive use of this fact for transferring results from polynomial to rational Krylov spaces.

It obviously suffices to consider *monic* polynomials q_{m-1} only. For computations it is convenient to have *nested* spaces $\mathcal{Q}_1 \subset \mathcal{Q}_2 \subset \cdots$, and such nested spaces are obtained if the polynomials q_{m-1} differ only by a linear factor as m is increased by one. For a given sequence of *poles* $\{\xi_1, \xi_2, \ldots\} \subset \overline{\mathbb{C}} \setminus \Lambda(A)$, $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$, we hence define

$$q_{m-1}(z) := \prod_{\substack{j=1 \\ \xi_j \neq \infty}}^{m-1} (z - \xi_j) \qquad \text{for } m = 1, 2, \ldots \tag{4.1}$$

once and for all. By convention, an empty product is equal to one and we will often refer to the polynomial q_{m-1} as "the poles of $\mathcal{Q}_m$," which also reflects the fact that the space $\mathcal{Q}_m$ is independent of the particular ordering of the poles. The resulting rational Krylov spaces are indeed nested.

Lemma 4.3. *Let q_{m-1} be defined by (4.1). If $M < \infty$ then*

$$\operatorname{span}\{\boldsymbol{b}\} = \mathcal{Q}_1 \subset \mathcal{Q}_2 \subset \cdots \subset \mathcal{Q}_M = \mathcal{K}_M,$$

otherwise

$$\operatorname{span}\{\boldsymbol{b}\} = \mathcal{Q}_1 \subset \mathcal{Q}_2 \subset \cdots$$

Proof. We first remark that $\mathcal{Q}_m \subseteq \mathcal{Q}_{m+1}$ since

$$\begin{aligned} \boldsymbol{v} \in \mathcal{Q}_m = q_{m-1}(A)^{-1}\mathcal{K}_m \quad &\Leftrightarrow \quad q_{m-1}(A)\boldsymbol{v} \in \mathcal{K}_m \\ &\Rightarrow \quad q_m(A)\boldsymbol{v} \in \mathcal{K}_{m+1} \\ &\Leftrightarrow \quad \boldsymbol{v} \in q_m(A)^{-1}\mathcal{K}_{m+1} = \mathcal{Q}_{m+1}. \end{aligned}$$

As long as $\dim(\mathcal{K}_{m+1}) = m+1$, we have $\mathcal{Q}_m \subset \mathcal{Q}_{m+1}$ by Lemma 4.2 (c). If $M < \infty$ then $\mathcal{K}_M$ is A-invariant, hence $q_{M-1}(A)p(A)\boldsymbol{b} \in \mathcal{K}_M$ and therefore $p(A)\boldsymbol{b} \in q_{M-1}(A)^{-1}\mathcal{K}_M = \mathcal{Q}_M$ for every polynomial p of arbitrary degree. Thus $\mathcal{Q}_M = \mathcal{K}_M$. □

Remark 4.4. Depending on the sequence $\{\xi_j\}$, various special cases of rational Krylov spaces exist.

(a) If all $\xi_j = \infty$ then $\mathcal{Q}_m = \mathcal{K}_m$ is a polynomial Krylov space.

(b) If all $\xi_j = \xi \in \mathbb{C}$ then $\mathcal{Q}_m$ is a *shift-and-invert Krylov space*, i.e.,

$$\mathcal{Q}_m(A, \boldsymbol{b}) = \mathcal{K}_m((A - \xi I)^{-1}, \boldsymbol{b}).$$

For the approximation of matrix functions such spaces were first considered in [MN04, EH06].

(c) If $\xi_{2j} = \infty$ and $\xi_{2j+1} = 0$ for all $j \geq 1$, one obtains the so-called *extended Krylov spaces* introduced in [DK98] and further studied in [JR09, KS10].

4.2 The Rayleigh–Ritz Method

A nice example of the close relationship between polynomial and rational Krylov spaces is the following lemma about Rayleigh quotients associated with $\mathcal{Q}_m$.

Lemma 4.5. *Let V_m be a basis of $\mathcal{Q}_m$, $A_m = V_m^\dagger A V_m$, and let χ_m denote the characteristic polynomial of A_m. Then the following statements hold.*

(a) *A_m is nonderogatory.*

(b) *$\chi_m(A) q_{m-1}(A)^{-1} \boldsymbol{b} \perp \mathcal{Q}_m(A, \boldsymbol{b})$.*

(c) *χ_m minimizes $\|s_m(A) q_{m-1}(A)^{-1} \boldsymbol{b}\|$ among all $s_m \in \mathcal{P}_m^\infty$.*

Proof. This lemma is obtained by simply replacing in Lemma 3.7 the vector $\boldsymbol{b}$ by $\boldsymbol{q} := q_{m-1}(A)^{-1}\boldsymbol{b}$ and using the fact that $\mathcal{K}_m(A, \boldsymbol{q}) = \mathcal{Q}_m(A, \boldsymbol{b})$. □

4.2.1 Rational Interpolation

The following lemma states that a Rayleigh approximation extracted from a rational Krylov space $\mathcal{Q}_m$ is exact for certain rational functions. This result follows from its polynomial counterpart and has been derived for special cases in several ways in the literature, e.g., in [DK98] for the extended Krylov spaces, or in [BR09].

Lemma 4.6. *Let V_m be a basis of $\mathcal{Q}_m$, $A_m = V_m^\dagger A V_m$, and let $P_m = V_m V_m^\dagger$ be the orthogonal projector onto $\mathcal{Q}_m$. Then for every rational function $\tilde{r}_m = p_m/q_{m-1} \in \mathcal{P}_m/q_{m-1}$ there holds*

$$P_m \tilde{r}_m(A)\boldsymbol{b} = V_m \tilde{r}_m(A_m) V_m^\dagger \boldsymbol{b}.$$

In particular, for every rational function $r_m = p_{m-1}/q_{m-1} \in \mathcal{P}_{m-1}/q_{m-1}$ there holds

$$r_m(A)\boldsymbol{b} = V_m r_m(A_m) V_m^\dagger \boldsymbol{b},$$

i.e., the Rayleigh approximation for $r_m(A)\boldsymbol{b}$ is exact (provided that $r_m(A_m)$ is defined).

Proof. Replacing in Lemma 3.9 the vector $\boldsymbol{b}$ by $\boldsymbol{q} := q_{m-1}(A)^{-1}\boldsymbol{b}$ yields

$$P_m p_m(A)\boldsymbol{q} = V_m p_m(A_m) V_m^\dagger \boldsymbol{q} \quad \text{for all } p_m \in \mathcal{P}_m. \tag{4.2}$$

Since $\boldsymbol{b} = q_{m-1}(A)\boldsymbol{q}$ and, again by Lemma 3.9, the Rayleigh approximation for $q_{m-1}(A)\boldsymbol{b}$ is exact, we have $\boldsymbol{b} = V_m q_{m-1}(A_m) V_m^\dagger \boldsymbol{q}$, or equivalently, $V_m^\dagger \boldsymbol{q} = q_{m-1}(A_m)^{-1} V_m^\dagger \boldsymbol{b}$. Replacing $V_m^\dagger \boldsymbol{q}$ in (4.2) yields the assertion. □

Remark 4.7. The eigenvalues $\Lambda(A_m)$ are called *rational Ritz values*. An example where $r_m(A)$ is defined but $r_m(A_m)$ is not can be obtained with the matrix $A = \operatorname{diag}(-2, -1, 1, 2)$, the vector $\boldsymbol{b} = [1, 1, 1, 1]^T$ and the poles $\xi_j = 0$ $(j = 1, 2, 3)$. By symmetry there is a rational Ritz value at zero when m is odd.

The following theorem states that the Rayleigh approximation $\boldsymbol{f}_m$ from a rational Krylov space is closely related to rational interpolation at the rational Ritz values $\Lambda(A_m)$. For this reason we will extend Definition 3.8 by using the name *Rayleigh–Ritz approximation* also in the case where the search space is a rational Krylov space.

Theorem 4.8. *Let V_m be a basis of $\mathcal{Q}_m$ and $A_m = V_m^\dagger A V_m$. Let f be a function such that $f(A_m)$ is defined. Then*

$$\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b} = r_m(A)\boldsymbol{b},$$

where $r_m = p_{m-1}/q_{m-1} \in \mathcal{P}_{m-1}/q_{m-1}$ interpolates f at $\Lambda(A_m)$.

Proof. With $\boldsymbol{q} := q_{m-1}(A)^{-1}\boldsymbol{b}$ and $\tilde{f} := f q_{m-1}$ we have $f(A)\boldsymbol{b} = \tilde{f}(A)\boldsymbol{q}$. By Lemma 3.10 the Rayleigh–Ritz approximation $\boldsymbol{f}_m$ for $\tilde{f}(A)\boldsymbol{q}$ from $\mathcal{K}_m(A, \boldsymbol{q}) = \mathcal{Q}_m(A, \boldsymbol{b})$ satisfies

$$\boldsymbol{f}_m = p_{m-1}(A)\boldsymbol{q} = p_{m-1}(A) q_{m-1}(A)^{-1}\boldsymbol{b},$$

where p_{m-1} interpolates $\tilde{f}$ at the nodes $\Lambda(A_m)$. Thus the function $r_m = p_{m-1}/q_{m-1}$ interpolates f at $\Lambda(A_m)$. □

We see from the proof of Theorem 4.8 that rational interpolation with a prescribed denominator q_{m-1} is in fact very similar to polynomial interpolation. As a consequence, there also exists a Cauchy integral representation of such rational interpolating functions. Let χ_m denote the characteristic polynomial of A_m and let Γ be an integration contour such that $\Lambda(A_m) \subset \operatorname{int}(\Gamma)$. If f is analytic in $\operatorname{int}(\Gamma)$ and extends continuously to Γ then so does $\tilde{f} = f q_{m-1}$. Owing to Hermite's formula [Wal69, p. 50], the polynomial $p_{m-1} \in \mathcal{P}_{m-1}$ interpolating $\tilde{f}$ at $\Lambda(A_m)$ can be expressed as

$$p_{m-1}(z) = \frac{1}{2\pi i} \int_\Gamma \frac{\chi_m(\zeta) - \chi_m(z)}{\chi_m(\zeta)(\zeta - z)} \tilde{f}(\zeta)\, \mathrm{d}\zeta.$$

For the interpolation error we have

$$\tilde{f}(z) - p_{m-1}(z) = \frac{1}{2\pi i} \int_\Gamma \frac{\chi_m(z)}{\chi_m(\zeta)(\zeta - z)} \tilde{f}(\zeta)\, \mathrm{d}\zeta.$$

Dividing this equation by q_{m-1} and setting $s_m := \chi_m / q_{m-1}$ we obtain

$$f(z) - r_m(z) = \frac{s_m(z)}{2\pi i} \int_\Gamma \frac{f(\zeta)}{s_m(\zeta)(\zeta - z)}\, \mathrm{d}\zeta,$$

where $r_m = p_{m-1}/q_{m-1}$ interpolates f at $\Lambda(A_m)$. By Theorem 4.8 we have $\boldsymbol{f}_m = r_m(A)\boldsymbol{b}$ and thus the error of the Rayleigh–Ritz approximation can be bounded as

$$\begin{aligned} \|f(A)\boldsymbol{b} - r_m(A)\boldsymbol{b}\| &\leq \frac{\ell(\Gamma)}{2\pi} \cdot \|s_m(A)\boldsymbol{b}\| \cdot \max_{\zeta\in\Gamma} \left\| \frac{f(\zeta)}{s_m(\zeta)} (\zeta I - A)^{-1} \right\| \\ &\leq D \cdot \|s_m(A)\boldsymbol{b}\| \cdot \max_{\zeta\in\Gamma} |s_m(\zeta)|^{-1}, \end{aligned} \tag{4.3}$$

where $\ell(\Gamma)$ denotes the length of Γ and $D = D(A, f, \Gamma)$ is a constant. It is remarkable that s_m is the minimizer of the factor $\|s_m(A)\boldsymbol{b}\|$ among all rational functions in $\mathcal{P}_m^\infty / q_{m-1}$ by Lemma 4.5 (c).

Here and in the following it will be essential for us to estimate $\|f(A)\|$. To this end let $\|f\|_\Sigma := \sup_{z\in\Sigma} |f(z)|$ denote the *uniform (or supremum) norm of* f on a set $\Sigma \supseteq \mathbb{W}(A)$ (we recall that $\mathbb{W}(A)$ denotes A's numerical range). Let us first assume that A is self-adjoint. Then by the spectral mapping theorem (Theorem 2.5) and the well-known fact that $\|A\| = \sup\{|z| : z \in \mathbb{W}(A)\}$ (cf. [Hal72, §24]) we have $\|f(A)\| \leq \|f\|_\Sigma$. In the more general case of a bounded operator A, the following theorem due to Crouzeix [Cro07] provides a powerful tool.

Theorem 4.9 (M. Crouzeix)**.** *Let f be analytic in a neighborhood of $\mathbb{W}(A)$, and let $\Sigma \supseteq \mathbb{W}(A)$. There holds $\|f(A)\| \leq C\|f\|_\Sigma$ with a constant $C \leq 11.08$. (It is conjectured that the bound also holds with $C = 2$.)*

If f is analytic in a neighborhood of $\mathbb{W}(A)$ containing an integration contour Γ such that $\mathbb{W}(A) \subseteq \Sigma \subset \mathrm{int}(\Gamma)$, then Theorem 4.9 applied to (4.3) yields

$$\|f(A)\boldsymbol{b} - r_m(A)\boldsymbol{b}\| \leq 2CD\|\boldsymbol{b}\| \cdot \|s_m\|_\Sigma \cdot \|s_m^{-1}\|_\Gamma,$$

noting that the zeros of s_m are rational Ritz values, hence contained in $\mathbb{W}(A)$, and therefore

$\|s_m^{-1}\|_\Gamma < \infty$. This bound suggests which type of approximation problems are connected with the optimization of a rational Krylov space for Rayleigh–Ritz extraction: consider rational functions s_m that are "smallest possible" on the set Σ and "largest possible" on the integration contour Γ winding around Σ. The zeros of the denominator q_{m-1} of such an optimal rational function should constitute "good" poles for the rational Krylov space $\mathcal{Q}_m$. These so-called *Zolotarev problems* will be treated in more detail in Chapter 7.

The derivation of error bounds using the Cauchy integral representation of the error is classical for studying the convergence of Padé(-type) approximations (cf. [Mag81, Eie84]). Indeed, if the poles q_{m-1} were free and we were given a set of $m-1$ additional interpolation nodes $\{\mu_1, \dots, \mu_{m-1}\}$ with nodal polynomial $\omega_{m-1} \in \mathcal{P}_{m-1}^{\infty}$, then it would also be possible to interpret the rational function $r_m = p_{m-1}/q_{m-1}$ from above as a *multi-point Padé-type approximation* (cf. [Sta96]) satisfying

$$\frac{q_{m-1}f - p_{m-1}}{\chi_m \omega_{m-1}} \quad \text{is bounded at each zero of } \chi_m\omega_{m-1}.$$

4.2.2 Near-Optimality

The accuracy of an approximation obtained by some approximation method is determined by the "quality" of the rational Krylov space $\mathcal{Q}_m$ and the extraction. Of course, an approximation $\boldsymbol{f}_m$ can only be as good as the search space it is extracted from, i.e.,

$$\min_{\boldsymbol{v} \in \mathcal{Q}_m} \|f(A)\boldsymbol{b} - \boldsymbol{v}\| \leq \|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|.$$

In a rational Krylov method it is therefore necessary to make the left-hand side of this inequality as small as possible by choosing the poles q_{m-1} suitably. If in addition the extraction is the Rayleigh–Ritz extraction, we will automatically obtain a near-best approximation $\boldsymbol{f}_m \in \mathcal{Q}_m$ (cf. [BR09, Prop. 3.1] for the polynomial Krylov case).

Theorem 4.10. *Let V_m be a basis of $\mathcal{Q}_m$ and $A_m = V_m^\dagger A V_m$. Let f be analytic in a neighborhood of $\mathbb{W}(A)$ and consider $\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b}$. For every set $\Sigma \supseteq \mathbb{W}(A)$ there holds*

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq 2C\|\boldsymbol{b}\| \min_{r_m \in \mathcal{P}_{m-1}/q_{m-1}} \|f - r_m\|_\Sigma.$$

with a constant $C \leq 11.08$. If A is self-adjoint the result holds with $C = 1$.

Proof. The Rayleigh–Ritz approximation is independent of the particular basis, hence let V_m be orthonormal. By Lemma 4.6 we know that $r_m(A)\boldsymbol{b} = V_m r_m(A_m)V_m^\dagger \boldsymbol{b}$ for every $r_m \in \mathcal{P}_{m-1}/q_{m-1}$. Thus,

$$\begin{aligned}\|f(A)\boldsymbol{b} - V_m f(A_m)V_m^\dagger \boldsymbol{b}\| &= \|f(A)\boldsymbol{b} - V_m f(A_m)V_m^\dagger \boldsymbol{b} - r_m(A)\boldsymbol{b} + V_m r_m(A_m)V_m^\dagger \boldsymbol{b}\| \\ &\leq \|\boldsymbol{b}\| \left(\|f(A) - r_m(A)\| + \|f(A_m) - r_m(A_m)\|\right) \\ &\leq 2C\|\boldsymbol{b}\| \cdot \|f - r_m\|_\Sigma,\end{aligned}$$

where we have used Theorem 4.9 for the last inequality. If A is self-adjoint then so is A_m and Theorem 4.9 holds with $C = 1$. The proof is completed by taking the infimum among all $r_m \in \mathcal{P}_{m-1}/q_{m-1}$ and the fact that this infimum is attained. □

The error bound of Theorem 4.10 will be an important tool for us, hence it deserves further discussion. First of all, we cannot expect this bound to be sharp if A is highly nonnormal since it is based on the numerical range $\mathbb{W}(A)$. Unfortunately, this bound can be crude even for a self-adjoint operator. To illustrate this we reconsider the model problem in the introduction, i.e., the solution of the 3D heat equation discretized by finite differences on the unit cube with $n = 15$ interior grid points in each Cartesian coordinate direction. This results in a symmetric negative definite matrix $A \in \mathbb{R}^{N\times N}$ with $N = 3375$ eigenvalues in $\mathbb{W}(A) \approx [-3043, -30]$. As in the introduction we approximate $f(A)\boldsymbol{b}$, where $f(z) = \exp(0.1z)$ and $\boldsymbol{b} \in \mathbb{R}^N$ is a random vector of unit length. In Figure 4.1 we show the absolute error curves of Rayleigh–Ritz approximations $\boldsymbol{f}_m$ from a polynomial and a rational Krylov space, respectively, and the errors of the orthogonal projection of $f(A)\boldsymbol{b}$ onto these spaces. The poles ξ_j of the rational Krylov space $\mathcal{Q}_m$ were all chosen equal to one. The dotted curves in Figure 4.1 are the error bounds[2] obtained from Theorem 4.10 in the polynomial and rational Krylov case, respectively. We observe that these bounds are not very close to the actual errors $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$. To explain this, we first note that for a symmetric (or Hermitian) matrix A the inequality in the proof of Theorem 4.10 can be improved to

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq 2\|\boldsymbol{b}\| \min_{r_m \in \mathcal{P}_{m-1}/q_{m-1}} \|f - r_m\|_{\Lambda(A)\cup\Lambda(A_m)},$$

[2]The associated best uniform approximation problems were solved approximately by the chebfun implementation of the Remez algorithm given in [PT09]. At least in the case of polynomial approximation the errors could also be bounded by the Chebyshev coefficients of the exponential function, which are explicitly known in terms of modified Bessel functions [Mei64, §6.5].

which is a min-max problem on the *discrete* set $\Lambda(A) \cup \Lambda(A_m)$. The error of the continuous minimizer is typically much larger than the error of the discrete minimizer, particularly if the eigenvalues $\Lambda(A)$ do not "fill" the interval $\mathbb{W}(A)$ sufficiently well. This is the reason why the bounds in Figure 4.1 are not good. In the worst case, for a symmetric matrix of size 2×2 having the extremal eigenvalues of A, we would obtain the same error bound as in Figure 4.1, although $\boldsymbol{f}_2$ would already be exact.

To restore the reputation of Theorem 4.10 we now let A be the discrete 1D Laplacian of (small) size 500×500, shifted and scaled so that $\mathbb{W}(A) \approx [-3043, -30]$, as above. Again we approximate $f(A)\boldsymbol{b}$ with $f(z) = \exp(0.1z)$ (which corresponds to a scaled solution of the 1D heat equation). The vector $\boldsymbol{b} \in \mathbb{R}^{500}$ of unit length is chosen as $\boldsymbol{b} = \sum_{j=1}^{500} \boldsymbol{u}_j / \sqrt{500}$ so that each of A's orthonormal eigenvectors $\boldsymbol{u}_j$ is active with equal weight. In Figure 4.2 we show the error curves of the polynomial and rational Rayleigh–Ritz approximations and the corresponding error bounds of Theorem 4.10. The error bounds are identical to the ones in Figure 4.1, but they are almost sharp for this problem. The reason is that the eigenvalues of the 1D Laplacian are a "very good discretization" of the interval $\mathbb{W}(A)$. Indeed, these eigenvalues are distributed according to the equilibrium measure of $\mathbb{W}(A)$, a notion we will make precise in Chapter 7. We expect that, at least for Hermitian matrices A, there should not be much room for improving the bound in Theorem 4.10 if only $\mathbb{W}(A)$ and $\|\boldsymbol{b}\|$ are taken into account.

Remark 4.11. Theorem 4.10 suggests that we choose the poles q_{m-1} such that

$$\min_{r_m \in \mathcal{P}_{m-1}/q_{m-1}} \|f - r_m\|_\Sigma$$

becomes as small as possible, which is a *rational best uniform approximation problem*. We will consider such problems in Chapter 7.

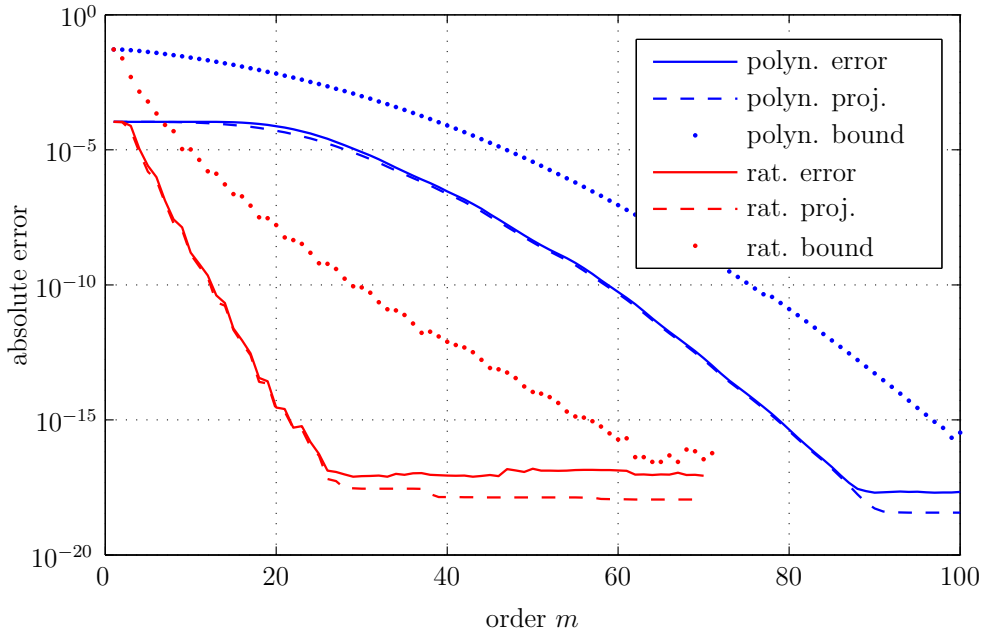


Figure 4.1: Convergence of Rayleigh–Ritz approximations from a polynomial (blue) and a rational Krylov space (red) for the solution of a 3D heat equation. The solid lines are the error curves $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$, and the dashed line is the error of the orthogonal projection of $f(A)\boldsymbol{b}$ onto the respective Krylov space. The dots indicate the error bounds obtained from Theorem 4.10.

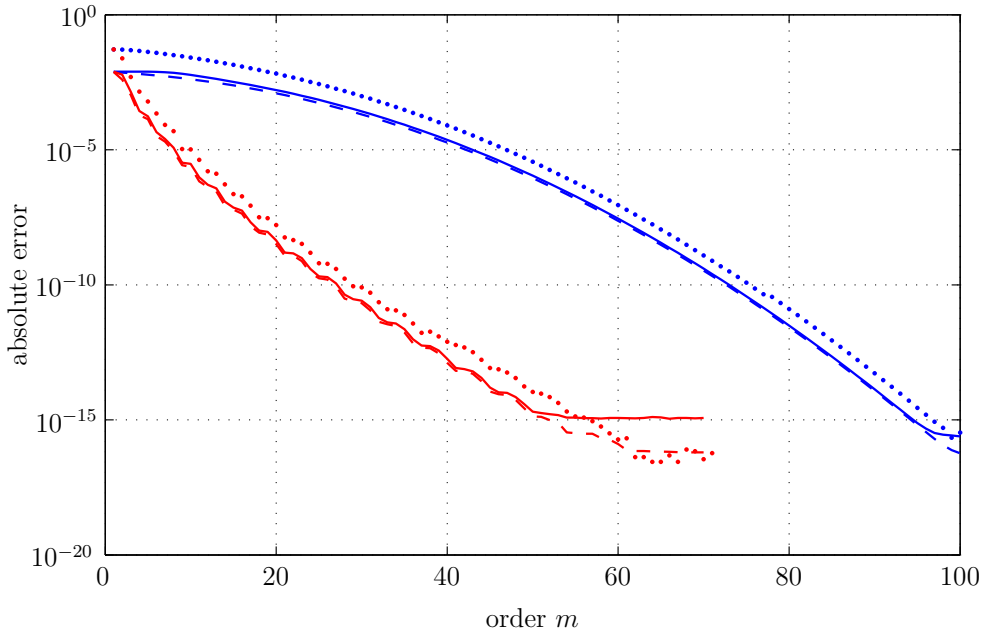


Figure 4.2: Convergence of Rayleigh–Ritz approximations for the solution of a 1D heat equation, where $\mathbb{W}(A)$ and $\|\boldsymbol{b}\|$ are the same as in the figure above, and hence the error bounds (dotted) are identical. However, for this problem these error bounds are far more satisfactory.

5 Rational Krylov Decompositions

In the previous chapter we presented Rayleigh–Ritz approximations as abstract approximations associated with a rational Krylov space $\mathcal{Q}_m$, independent of the basis V_m. However, each computational method requires a basis of the vector space it is working with. Thus when it comes to actually computing a Rayleigh–Ritz approximation, the basis V_m eventually needs to be constructed. In 1984, A. Ruhe [Ruh84] introduced the *rational Arnoldi algorithm*[1] to iteratively compute an ascending orthonormal basis of a rational Krylov space. It can be thought of as an extension of the spectrally transformed Arnoldi algorithm, allowing the pole $\xi = \xi_j$ to vary in each iteration. Ten years later, Ruhe reconsidered the rational Arnoldi algorithm for generalized eigenproblems $A\boldsymbol{x} = \lambda B\boldsymbol{x}$, see [Ruh94b, Ruh94c, Ruh94a] and also [Ruh98]. He showed that the basis vectors generated by his algorithm satisfy a *rational Arnoldi decomposition*. In our context these decompositions are essential for the efficient and stable computation of Rayleigh–Ritz approximations for $f(A)\boldsymbol{b}$.

In Section 5.1 we review the rational Arnoldi algorithm and rational Arnoldi decompositions generated therein. The relationship between these decompositions and orthogonal rational functions is discussed in Section 5.2. In Section 5.3 we introduce rational Krylov decompositions, a general framework that allows us to relate existing rational Krylov methods and to derive some novel rational Krylov algorithms for the approximation of operator functions in Section 5.4.

[1] Ruhe actually used the name *rational Krylov sequence algorithm*, but for reasons which become apparent in Section 5.3 we prefer to use *rational Arnoldi algorithm* here.

5.1 The Rational Arnoldi Algorithm

Before describing the rational Arnoldi algorithm we first show that it is always possible to iteratively construct a basis for a rational Krylov space $\mathcal{Q}_{j+1}$ given a basis for $\mathcal{Q}_j$. As in the previous chapters, M denotes the invariance index of the Krylov space.

Lemma 5.1. *There exists a vector $\boldsymbol{y}_j \in \mathcal{Q}_j$ such that $(I - A/\xi_j)^{-1} A\boldsymbol{y}_j \in \mathcal{Q}_{j+1} \setminus \mathcal{Q}_j$ if and only if $j < M$.*

Proof. Set $\boldsymbol{q} := q_{j-1}(A)^{-1}\boldsymbol{b}$. By the definition of a rational Krylov space we have

$$(I - A/\xi_j)^{-1} A\mathcal{Q}_j = (I - A/\xi_j)^{-1} A\mathcal{K}_j(A, \boldsymbol{q}) = A\mathcal{K}_j(A, q_j(A)^{-1}\boldsymbol{b}) \subseteq \mathcal{Q}_{j+1}.$$

Let $j < M$ and assume there exists no vector $\boldsymbol{y}_j$ such that $(I - A/\xi_j)^{-1} A\boldsymbol{y}_j \in \mathcal{Q}_{j+1} \setminus \mathcal{Q}_j$. Then $(I - A/\xi_j)^{-1} A\mathcal{Q}_j \subseteq \mathcal{Q}_j$ and hence $A\mathcal{Q}_j \subseteq (I - A/\xi_j)\mathcal{Q}_j$. Since $\mathcal{Q}_j$ is not A-invariant, $\dim(A\mathcal{Q}_j) = \dim((I - A/\xi_j)\mathcal{Q}_j) = j$ and therefore $A\mathcal{Q}_j = (I - A/\xi_j)\mathcal{Q}_j$. This is equivalent to $A\mathcal{K}_j(A, \boldsymbol{q}) = (I - A/\xi_j)\mathcal{K}_j(A, \boldsymbol{q})$, which is a contradiction because $(I - A/\xi_j)\boldsymbol{q} \notin A\mathcal{K}_j(A, \boldsymbol{q})$.

For $j \geq M$ we have $(I - A/\xi_j)^{-1} A\mathcal{Q}_j = \mathcal{Q}_j$ and the rational Krylov space $\mathcal{Q}_j$ cannot be enlarged further. □

Let $\{\xi_1, \xi_2, \ldots\} \subset \overline{\mathbb{C}} \setminus \Lambda(A)$ be a given sequence of *nonzero* poles. The rational Arnoldi algorithm now proceeds as follows:

In the first iteration $j = 1$ we set $\boldsymbol{v}_1 = \boldsymbol{b}/\|\boldsymbol{b}\|$, which is an orthonormal basis vector of $\mathcal{Q}_1$. In the following iterations the vector $\boldsymbol{v}_{j+1}$ is obtained by orthonormalizing

$$\boldsymbol{x}_j = (I - A/\xi_j)^{-1} A\boldsymbol{y}_j, \quad \boldsymbol{y}_j = \sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j} \tag{5.1}$$

against the already known orthonormal vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j$. The vector $\boldsymbol{y}_j \in \mathcal{Q}_j$ is called a *continuation vector* and it is chosen such that $\boldsymbol{x}_j$ is not a linear combination of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j$. By Lemma 5.1 this is possible if and only if $j < M$, i.e., as long as we have not reached the

invariance index of the rational Krylov space. The orthogonalization leads to the equation

$$\boldsymbol{x}_j = \sum_{i=1}^{j+1} \boldsymbol{v}_i h_{i,j}, \tag{5.2}$$

where the normalization coefficient $h_{j+1,j}$ can be chosen > 0. At this point we have computed an orthonormal basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{j+1}\}$ of $\mathcal{Q}_{j+1}$.

Equating (5.1) and (5.2) and left-multiplying both sides by $I - A/\xi_j$ gives

$$A \sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j} = (I - A/\xi_j) \sum_{i=1}^{j+1} \boldsymbol{v}_i h_{i,j}, \tag{5.3}$$

and separation of the terms containing A yields

$$A \left(\sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j} + \sum_{i=1}^{j+1} \boldsymbol{v}_i h_{i,j} \xi_j^{-1} \right) = \sum_{i=1}^{j+1} \boldsymbol{v}_i h_{i,j}. \tag{5.4}$$

For $j = 1, \ldots, m < M$ we can rewrite (5.4) in "matrix language" as a *rational Arnoldi decomposition*

$$AV_m(U_m + H_m D_m) + A\boldsymbol{v}_{m+1} h_{m+1,m} \xi_m^{-1} \boldsymbol{e}_m^T = V_m H_m + \boldsymbol{v}_{m+1} h_{m+1,m} \boldsymbol{e}_m^T, \tag{5.5}$$

where

$$\begin{aligned}
V_m &= [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m] \quad \text{is an orthonormal basis of } \mathcal{Q}_m, \\
[V_m, \boldsymbol{v}_{m+1}] &\quad \text{is an orthonormal basis of } \mathcal{Q}_{m+1}, \\
H_m &= [h_{i,j}] \in \mathbb{C}^{m \times m} \quad \text{is an unreduced upper Hessenberg matrix}, \\
h_{m+1,m} &\quad \text{is positive}, \\
U_m &= [u_{i,j}] \in \mathbb{C}^{m \times m} \quad \text{is an upper triangular matrix}, \\
D_m &= \operatorname{diag}(\xi_1^{-1}, \ldots, \xi_m^{-1}), \\
\boldsymbol{e}_m &\quad \text{denotes the } m\text{th unit coordinate vector in } \mathbb{R}^m.
\end{aligned}$$

Setting

$$\underline{H_m} := \begin{bmatrix} H_m \\ h_{m+1,m} \boldsymbol{e}_m^T \end{bmatrix} \quad \text{and} \quad \underline{K_m} := \begin{bmatrix} U_m + H_m D_m \\ h_{m+1,m} \xi_m^{-1} \boldsymbol{e}_m^T \end{bmatrix}, \tag{5.6}$$

one can write (5.5) more succinctly as

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m}, \tag{5.7}$$

where $\underline{H_m}$ and $\underline{K_m}$ are unreduced upper Hessenberg matrices of size $(m+1) \times m$ (the underline is intended to symbolize the additional last row).

Note that if $U_m = I_m$ and all poles ξ_j are infinite then $D_m = O$ and (5.7) reduces to a *polynomial Arnoldi decomposition*

$$AV_m = V_{m+1}\underline{H_m}.$$

In this case Algorithm 1 reduces to the polynomial Arnoldi algorithm.

Algorithm 1: Rational Arnoldi algorithm.

Input: $A, \boldsymbol{b}, \{\xi_1, \ldots, \xi_m\}, U_m$

$\boldsymbol{v}_1 := \boldsymbol{b}/\|\boldsymbol{b}\|$
for $j = 1, \ldots, m$ **do**
  $\boldsymbol{y} := \sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j}$
  $\boldsymbol{x} := (I - A/\xi_j)^{-1} A\boldsymbol{y}$
  for $i = 1, \ldots, j$ **do**
    $h_{i,j} := \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle$
    $\boldsymbol{x} := \boldsymbol{x} - \boldsymbol{v}_i h_{i,j}$
  $h_{j+1,j} := \|\boldsymbol{x}\|$
  $\boldsymbol{v}_{j+1} := \boldsymbol{x}/h_{j+1,j}$

Remark 5.2. Our construction does not allow for poles at zero. However, this is no practical restriction since we can choose a number $\sigma \in \mathbb{C}$ different from all the poles and run the rational Arnoldi algorithm with the shifted operator $A - \sigma I$ and nonzero shifted poles $\xi_j - \sigma$ as input parameters. It is also possible to vary the shift $\sigma = \sigma_j$ in each iteration, and if these shifts are chosen properly, this can improve the robustness of the rational Arnoldi algorithm for eigenvalue computations [LM98].

Remark 5.3. The rational Krylov space can be enlarged differently than is done in (5.1). As an example, assume that all the poles ξ_j are finite. Analogously to Lemma 5.1 one can

easily verify that there exists a *continuation vector* $\boldsymbol{y}_j \in \mathcal{Q}_j$ such that $(A - \xi_j I)^{-1}\boldsymbol{y}_j \in \mathcal{Q}_{j+1} \setminus \mathcal{Q}_j$ if and only if $j < M$. Hence we can derive a variant of the rational Arnoldi algorithm where the vector $\boldsymbol{v}_{j+1}$ is obtained by orthonormalizing

$$\boldsymbol{x}_j = (A - \xi_j I)^{-1}\boldsymbol{y}_j, \quad \boldsymbol{y}_j = \sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j},$$

against the orthonormal vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j$, thus yielding an orthonormal basis of $\mathcal{Q}_{j+1}$. The resulting rational Arnoldi decomposition is

$$AV_{m+1}\underline{H_m} = V_{m+1}(\underline{U_m} + \underline{H_m}X_m), \tag{5.8}$$

where (for $m < M$)

$$\begin{aligned} V_{m+1} &= [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m+1}] \quad \text{is an orthonormal basis of } \mathcal{Q}_{m+1}, \\ \underline{H_m} &= [h_{i,j}] \in \mathbb{C}^{(m+1)\times m} \quad \text{is an unreduced upper Hessenberg matrix}, \\ \underline{U_m} &= [u_{i,j}] \in \mathbb{C}^{(m+1)\times m} \quad \text{is an upper triangular matrix}, \\ X_m &= \operatorname{diag}(\xi_1, \ldots, \xi_m). \end{aligned}$$

5.2 Orthogonal Rational Functions

Let A be a bounded *normal* operator on the Hilbert space $\mathcal{H}$, i.e., $A^*A = AA^*$.

We will require some notions from spectral theory. A *spectral measure* E is a function defined on the Borel sets of $\mathbb{C}$ whose values are orthogonal projection operators defined on $\mathcal{H}$ such that $E(\mathbb{C}) = I$ and $E(\bigcup_j \Omega_j) = \sum_j E(\Omega_j)$ for every sequence $\{\Omega_j\}$ of disjoint Borel sets. Let φ be a function such that $\varphi(A)$ is defined. By the *spectral theorem for normal operators* (cf. [Hal72, Thm. 44.1]) there exists a unique spectral measure E with compact support such that $\langle \varphi(A)\boldsymbol{x}, \boldsymbol{y}\rangle = \int \varphi(\lambda)\,\mathrm{d}\langle E(\lambda)\boldsymbol{x}, \boldsymbol{y}\rangle$ for every pair of vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$. The dependence of $\varphi(A)$ on φ and E will be denoted by $\varphi(A) = \int \varphi(\lambda)\,\mathrm{d}E(\lambda)$.

Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m+1}\}$ be an orthonormal basis of $\mathcal{Q}_{m+1}$ and let $\{\varphi_1, \ldots, \varphi_{m+1}\} \subset \mathcal{P}_m/q_m$ be the unique rational functions such that $\boldsymbol{v}_j = \varphi_j(A)\boldsymbol{b}$ for $j = 1, \ldots, m+1$ (the uniqueness following from Lemma 4.2 (d)). Owing to the usual calculus of spectral integrals

(cf. [Hal72, §36–37]), there holds

$$\begin{aligned}\langle \boldsymbol{v}_k, \boldsymbol{v}_j \rangle &= \langle \varphi_k(A)\boldsymbol{b}, \varphi_j(A)\boldsymbol{b} \rangle \\ &= \big\langle \textstyle\int \varphi_k(\lambda)\,\mathrm{d}E(\lambda)\boldsymbol{b}, \int \varphi_j(\lambda)\,\mathrm{d}E(\lambda)\boldsymbol{b} \big\rangle \\ &= \int \varphi_k(\lambda)\overline{\varphi_j(\lambda)}\,\mathrm{d}\langle E(\lambda)\boldsymbol{b}, E(\lambda)\boldsymbol{b}\rangle \\ &=: \langle \varphi_k, \varphi_j \rangle_E,\end{aligned}$$

where we have used the fact that $\Omega_1 \cap \Omega_2 = \emptyset$ implies $E(\Omega_1)E(\Omega_2) = O$ (spectral measures are *multiplicative*, see [Hal72, Thm. 36.2]) to reduce the double integral. We have thus established that $\{\varphi_1, \ldots, \varphi_{m+1}\}$ constitutes an orthonormal basis of the Hilbert space $\mathcal{P}_m/q_m$ with inner product $\langle \cdot, \cdot \rangle_E$, being isometric isomorphic to $\mathcal{Q}_{m+1}$.

Let us now consider the special case where all poles satisfy $\xi_j \in \mathbb{R} \cup \{\infty\}$ and A is self-adjoint. The spectral measure E is then supported on a subset of the real line (cf. [Hal72, Thm. 43.1]). From the theory of orthogonal rational functions (cf. [BGHN99, Thm. 11.1.2] and [BGHN03]) we know that it should be possible to construct a set $\{\varphi_1, \ldots, \varphi_{m+1}\}$ of rational functions orthonormal with respect to $\langle \cdot, \cdot \rangle_E$ by a three-term recurrence.[2] The construction of such a recursion in rational Krylov spaces was given in [DB07]. Here is a simple derivation: we go back to the rational Arnoldi algorithm in Section 5.1 and assume that the continuation vector $\boldsymbol{y}_j$ can always be chosen as the last computed basis vector $\boldsymbol{v}_j$ (for $j = 1, \ldots, m$). In this case, (5.3) can be written in terms of rational functions

$$\varphi_{j+1}(z)h_{j+1,j} = \frac{z}{1 - z/\xi_j}\varphi_j(z) - \varphi_j(z)h_{j,j} - \cdots - \varphi_1(z)h_{1,j}, \tag{5.9}$$

with the initial condition $\varphi_1 \equiv \langle 1, 1 \rangle_E^{1/2}$.

Defining $W_{m+1} := [\varphi_1, \ldots, \varphi_{m+1}]$, the unreduced upper Hessenberg matrix $\underline{H_m} := [h_{i,j}] \in \mathbb{C}^{(m+1)\times m}$ and the diagonal matrix $D_m := \mathrm{diag}(\xi_1^{-1}, \ldots, \xi_m^{-1})$, (5.9) can be rewritten as

$$zW_{m+1}(\underline{I_m} + \underline{H_m}D_m) = W_{m+1}\underline{H_m}, \tag{5.10}$$

which is nothing but the "scalarized" form of the rational Arnoldi decomposition (5.7) (see

[2] This also follows from the well-known fact that such a recurrence exists for orthonormal polynomials $\{\pi_1, \ldots, \pi_{m+1}\}$ with respect to the inner product $\langle \pi_k, \pi_j \rangle_{E_{\boldsymbol{q}}} := \int \pi_k(\lambda)\overline{\pi_j(\lambda)}\,\mathrm{d}\langle E(\lambda)\boldsymbol{q}, E(\lambda)\boldsymbol{q}\rangle$ with $\boldsymbol{q} := q_m(A)^{-1}\boldsymbol{b}$.

also (5.6)). Note that the coefficients $h_{i,j}$ are coordinates of $\varphi_{j+1}h_{j+1,j}$ in the (nonorthogonal) basis

$$\Big\{\frac{z}{1-z/\xi_j}\varphi_j(z), \varphi_j(z), \dots, \varphi_1(z)\Big\}.$$

Of course, we could also represent φ_{j+1} in some other basis, for example,

$$\varphi_{j+1}(z)\tilde{h}_{j+1,j} = \frac{z}{1-z/\xi_j}\varphi_j(z) - \tilde{\varphi}_j(z)\tilde{h}_{j,j} - \cdots - \tilde{\varphi}_1(z)\tilde{h}_{1,j}, \tag{5.11}$$

where

$$\tilde{\varphi}_i(z) := \frac{1-z/\xi_{i-1}}{1-z/\xi_j}\varphi_i(z) \quad \text{for } i = 1, \dots, j,$$

$\tilde{h}_{j+1,j}$ is positive and $\xi_0 := \infty$, for convenience. Multiplying (5.11) by $1 - z/\xi_j$ and separating terms containing the factor z yields

$$z\Big(\varphi_{j+1}\frac{\tilde{h}_{j+1,j}}{\xi_j} + \varphi_j + \varphi_j\frac{\tilde{h}_{j,j}}{\xi_{j-1}} + \cdots + \varphi_1\frac{\tilde{h}_{1,j}}{\xi_0}\Big) = \varphi_{j+1}\tilde{h}_{j+1,j} + \cdots + \varphi_1\tilde{h}_{1,j}.$$

For $j = 1, \dots, m$ this can be rewritten as

$$zW_{m+1}(\underline{I_m} + \widetilde{D}_m\underline{\widetilde{H}_m}) = W_{m+1}\underline{\widetilde{H}_m}, \tag{5.12}$$

where $\underline{\widetilde{H}_m} := [\tilde{h}_{i,j}] \in \mathbb{C}^{(m+1)\times m}$ is an unreduced upper Hessenberg matrix and $\widetilde{D}_m := \operatorname{diag}(\xi_0^{-1}, \xi_1^{-1}, \dots, \xi_m^{-1})$. Note that (5.12) and (5.10) look very similar, except that the order of the factors "D" and "H" is switched. It is remarkable that this new order makes $\widetilde{H}_m = [I_m, \boldsymbol{0}]\underline{\widetilde{H}_m}$ symmetric and therefore tridiagonal: by (5.12) we have

$$\underline{I_m} + \widetilde{D}_m\underline{\widetilde{H}_m} = W_{m+1}^* z^{-1} W_{m+1}\underline{\widetilde{H}_m},$$

and right-multiplying by $\widetilde{H}_m^{-1}$ yields[3]

$$\underline{\widetilde{H}_m^{-1}} + \widetilde{D}_m\underline{\widetilde{I}_m} = W_{m+1}^* z^{-1} W_{m+1}\underline{\widetilde{I}_m},$$

where $\underline{\widetilde{I}_m}$ and $\underline{\widetilde{H}_m^{-1}}$ are I_m and $\widetilde{H}_m^{-1}$ appended with a nonzero row at the bottom, respectively. Since $W_m^* z^{-1} W_m$ is Hermitian for all $z \in \mathbb{R}$, we know that $\widetilde{H}_m^{-1}$ is Hermitian

[3] If $\widetilde{H}_m$ is not invertible, consider (5.12) for the shifted variable $z - \sigma$ instead of z. For more properties of $\widetilde{H}_m$ and its relation to H_m we refer to [Fas05].

and hence $\widetilde{H}_m$ must be symmetric (all entries $\tilde{h}_{j+1,j}$ are positive) and thus tridiagonal. In other words, the orthogonal rational functions φ_j satisfy a three-term recurrence in (5.12).

To get a nicer notation we define

$$\alpha_j := \tilde{h}_{j,j} \quad \text{and} \quad \beta_j := \tilde{h}_{j+1,j} = \tilde{h}_{j,j+1},$$

such that (5.11) becomes

$$\beta_j \varphi_{j+1}(z) = \frac{z}{1 - z/\xi_j}\varphi_j(z) - \alpha_j \frac{1 - z/\xi_{j-1}}{1 - z/\xi_j}\varphi_j(z) - \beta_{j-1}\frac{1 - z/\xi_{j-2}}{1 - z/\xi_j}\varphi_{j-1}(z). \tag{5.13}$$

To reduce operations with z we follow [DB07] by introducing

$$r_j(z) = \frac{z[\varphi_j(z) + \beta_{j-1}\xi_{j-2}^{-1}\varphi_{j-1}(z)] - \beta_{j-1}\varphi_{j-1}(z)}{1 - z/\xi_j} \quad \text{and} \quad s_j(z) = \frac{1 - z/\xi_{j-1}}{1 - z/\xi_j}\varphi_j(z),$$

such that (5.13) is equivalent to

$$\beta_j \varphi_{j+1}(z) = r_j(z) - \alpha_j s_j(z), \tag{5.14}$$

a formula from which α_j and β_j can be easily computed as

$$\alpha_j = \frac{\langle r_j, \varphi_j \rangle_E}{\langle s_j, \varphi_j \rangle_E} \quad \text{and} \quad \beta_j = \|r_j(z) - \alpha_j s_j(z)\|_E.$$

Here, $\|\cdot\|_E$ denotes the norm induced by the inner product $\langle \cdot, \cdot \rangle_E$. The *rational Lanczos algorithm* is obtained by replacing z by A in (5.14) and right-multiplying the result by $\boldsymbol{b}$, see Algorithm 2. This algorithm reduces to the polynomial Lanczos algorithm if all poles ξ_j are at infinity. Note that in each iteration of Algorithm 2 *two* linear systems with $I - A/\xi_j$ need to be solved (except if $\xi_{j-1} = \xi_j$ because then $s_j = \varphi_j$). Hence, this algorithm is in general not competitive with the rational Arnoldi algorithm if the poles ξ_j vary often. Moreover, we will make explicit use of the orthogonality of the rational Krylov basis V_{m+1} when computing Rayleigh–Ritz approximations for $f(A)\boldsymbol{b}$ (see Chapter 6). In this case full orthogonalization of V_{m+1} is required anyway and one cannot take advantage of the short recurrence.

Remark 5.4. We have tacitly assumed that a representation (5.11) exists. This assumption may fail, e.g., if φ_j happens to have a zero at ξ_j. This situation is often called

unlucky breakdown: we are not able to compute a basis of $\mathcal{Q}_{j+1}$ although this space has dimension $j+1$. Rational functions for which such breakdowns do not occur are called *regular* [BGHN99, Ch. 11].

Algorithm 2: Rational Lanczos algorithm.

Input: $A, \boldsymbol{b}, \{\xi_1, \ldots, \xi_m\}$

$\xi_{-1} := \infty$

$\xi_0 := \infty;\ \beta_0 := 0;\ \boldsymbol{v}_0 := \boldsymbol{0}$

$\boldsymbol{v}_1 := \boldsymbol{b}/\|\boldsymbol{b}\|$

for $j = 1, \ldots, m$ **do**

$\widetilde{\boldsymbol{y}} := A(\boldsymbol{v}_j + \beta_{j-1}\xi_{j-2}^{-1}\boldsymbol{v}_{j-1}) - \beta_{j-1}\boldsymbol{v}_{j-1}$

$\boldsymbol{r} := (I - A/\xi_j)^{-1}\widetilde{\boldsymbol{y}}$

$\boldsymbol{s} := (I - A/\xi_j)^{-1}(I - A/\xi_{j-1})\boldsymbol{v}_j$

$\alpha_j := (\boldsymbol{r}, \boldsymbol{v}_j)/(\boldsymbol{s}, \boldsymbol{v}_j)$

$\boldsymbol{w} := \boldsymbol{r} - \alpha_j \boldsymbol{s}$

$\beta_j := \|\boldsymbol{w}\|$

$\boldsymbol{v}_{j+1} := \boldsymbol{w}/\|\boldsymbol{w}\|$

5.3 Rational Krylov Decompositions

We have seen that the rational Arnoldi algorithm or variants of it lead to decompositions of the form (5.7), (5.8) or (5.12). We will find it fruitful to introduce the following generalization of such decompositions.

Definition 5.5. A relation

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m}, \tag{5.15}$$

where $V_{m+1} = [V_m, \boldsymbol{v}_{m+1}]$ has $m+1$ linearly independent columns such that $\mathcal{R}(V_{m+1}) = \mathcal{Q}_{m+1}$, $\mathcal{R}(V_m) = \mathcal{Q}_m$, $\underline{K_m} \in \mathbb{C}^{(m+1)\times m}$, $\underline{H_m} \in \mathbb{C}^{(m+1)\times m}$, and $\underline{H_m}$ is of rank m, is called a *rational Krylov decomposition*.

If the last row of $\underline{K_m}$ contains only zeros we have

$$AV_mK_m = V_{m+1}\underline{H_m} \tag{5.16}$$

and say that this decomposition is *reduced*.

Let us collect some useful facts about rational Krylov decompositions.

Lemma 5.6.

(a) *The matrix* $\underline{K_m}$ *of* (5.15) *is of rank* m*. In particular, the matrix* K_m *of the reduced rational Krylov decomposition* (5.16) *is invertible.*

(b) *The validity of* (5.16) *implies* $\boldsymbol{v}_{m+1} \in A\mathcal{Q}_m \setminus \mathcal{Q}_m$*.*

(c) *If* (5.16) *is an orthonormal decomposition, i.e.,* $V_{m+1}^* V_{m+1} = I_{m+1}$*, then the Rayleigh quotient* $A_m = V_m^* A V_m$ *can be computed as* $A_m = H_m K_m^{-1}$*.*

Proof.

(a) Since the right-hand side of (5.15) is of rank m, the same must hold for the left-hand side. We have $m = \operatorname{rank}(AV_{m+1}\underline{K_m}) \leq \min\{\operatorname{rank}(AV_{m+1}), \operatorname{rank}(\underline{K_m})\}$, hence $\operatorname{rank}(\underline{K_m}) = m$.

(b) The decomposition (5.16) can be written as $AV_mK_m = V_mH_m + \boldsymbol{v}_{m+1}\boldsymbol{h}_m^T$, where $\boldsymbol{h}_m^T \in \mathbb{C}^{1\times m}$ denotes the last row of $\underline{H_m}$. Therefore $\boldsymbol{v}_{m+1}$ is a linear combination of vectors in $\mathcal{R}(AV_m) = A\mathcal{Q}_m$ and $\mathcal{R}(V_m) = \mathcal{Q}_m$. However, $\boldsymbol{v}_{m+1}$ cannot be contained in $\mathcal{Q}_m$ since by definition of a rational Krylov decomposition $\mathcal{R}([V_m, \boldsymbol{v}_{m+1}]) = \mathcal{Q}_{m+1}$.

(c) There holds $V_m^* V_{m+1} = [I_m, \boldsymbol{0}]$, and hence $V_m^* A V_m K_m = V_m^* V_{m+1} \underline{H_m} = H_m$. □

Remark 5.7. The assumption that $\underline{H_m}$ is of full rank m is satisfied if the decomposition is generated by the rational Arnoldi algorithm (or one of its variants), since $\underline{H_m}$ is unreduced upper Hessenberg in this case.

We define the *rational Krylov approximation*

$$\boldsymbol{f}_m^{\mathrm{RK}} := V_m f(H_m K_m^{-1}) V_m^\dagger \boldsymbol{b} \tag{5.17}$$

associated with the rational Krylov decompositions (5.15) or (5.16), provided that the matrix function $f(H_mK_m^{-1})$ is defined. Of particular interest are approximations associated with reduced decompositions.

Theorem 5.8. *The rational Krylov approximation associated with the reduced rational Krylov decomposition* (5.16) *satisfies*

$$\boldsymbol{f}_m^{\mathrm{RK}} = V_m f(H_m K_m^{-1}) V_m^\dagger \boldsymbol{b} = r_m(A)\boldsymbol{b},$$

where $r_m \in \mathcal{P}_{m-1}/q_{m-1}$ *interpolates* f *at the eigenvalues* $\Lambda(H_m K_m^{-1})$.

Proof. By Lemma 5.6 (a), K_m is invertible and the decomposition (5.16) can be written as a polynomial Krylov decomposition

$$AV_m = V_{m+1}\underline{H_m}K_m^{-1} = V_m H_m K_m^{-1} + \boldsymbol{v}_{m+1}\boldsymbol{h}_m^T K_m^{-1}, \tag{5.18}$$

where $\boldsymbol{h}_m^T$ denotes the last row of $\underline{H_m}$ and $\boldsymbol{v}_{m+1} \in \mathcal{Q}_{m+1} \setminus \mathcal{Q}_m$. By Lemma 5.6 (b) we know that $\boldsymbol{v}_{m+1} \in A\mathcal{Q}_m$. Therefore $\mathcal{Q}_m = \mathcal{K}_m(A, \boldsymbol{q})$ and $\mathcal{Q}_{m+1} = \mathcal{K}_{m+1}(A, \boldsymbol{q})$ with $\boldsymbol{q} := q_{m-1}(A)^{-1}\boldsymbol{b}$.

By induction we show that $A^j\boldsymbol{q} = V_m(H_m K_m^{-1})^j V_m^\dagger \boldsymbol{q}$ for all $j \le m-1$, which is obviously true for $j = 0$. Assume that the assertion is true for some $j \le m-2$. Then

$$\begin{aligned} A^{j+1}\boldsymbol{q} &= AA^j\boldsymbol{q} = AV_m(H_m K_m^{-1})^j V_m^\dagger \boldsymbol{q} \\ &= (V_m H_m K_m^{-1} + \boldsymbol{v}_{m+1}\boldsymbol{h}_m^T K_m^{-1})(H_m K_m^{-1})^j V_m^\dagger \boldsymbol{q} \\ &= V_m(H_m K_m^{-1})^{j+1} V_m^\dagger \boldsymbol{q}, \end{aligned}$$

where we have used (5.18) and the facts that $A^{j+1}\boldsymbol{q} \in \mathcal{Q}_m$ and $\boldsymbol{v}_{m+1} \in \mathcal{Q}_{m+1} \setminus \mathcal{Q}_m$, hence the coefficient $\boldsymbol{h}_m^T K_m^{-1}(H_m K_m^{-1})^j V_m^\dagger \boldsymbol{q}$ must vanish. We have thus established that $p_{m-1}(A)\boldsymbol{q} = V_m p_{m-1}(H_m K_m^{-1}) V_m^\dagger \boldsymbol{q}$ for all $p_{m-1} \in \mathcal{P}_{m-1}$. Using this relation we obtain

$$\begin{aligned} V_m f(H_m K_m^{-1}) V_m^\dagger \boldsymbol{b} &= V_m f(H_m K_m^{-1}) V_m^\dagger (q_{m-1}(A)\boldsymbol{q}) \\ &= V_m f(H_m K_m^{-1}) q_{m-1}(H_m K_m^{-1}) V_m^\dagger \boldsymbol{q} \\ &= V_m p_{m-1}(H_m K_m^{-1}) V_m^\dagger \boldsymbol{q} = p_{m-1}(A)\boldsymbol{q}, \end{aligned}$$

where p_{m-1} interpolates $\tilde{f} := f q_{m-1}$ at the nodes $\Lambda(H_m K_m^{-1})$. Thus the function $r_m = p_{m-1}/q_{m-1}$ interpolates f at $\Lambda(H_m K_m^{-1})$. □

5.4 Various Rational Krylov Methods

5.4.1 A Restarted Rational Krylov Method

Let us consider a reduced rational Krylov decomposition,

$$AV^{(1)}K^{(1)} = V^{(1)}H^{(1)} + \boldsymbol{v}^{(1)}\boldsymbol{h}^{(1)}, \tag{5.19}$$

being computed, e.g., by m iterations of the rational Arnoldi algorithm. More precisely, $\mathcal{R}(V^{(1)}) = \mathcal{Q}_m$, $\mathcal{R}([V^{(1)}, \boldsymbol{v}^{(1)}]) = \mathcal{Q}_{m+1}$, and $K^{(1)}, H^{(1)}$ are $m \times m$ matrices, $\boldsymbol{h}^{(1)} \in \mathbb{C}^{1\times m}$ is a *row vector* (for notational convenience). We may also assume that $\|\boldsymbol{b}\|V^{(1)}\boldsymbol{e}_1 = \boldsymbol{b}$. The rational Krylov approximation for $f(A)\boldsymbol{b}$ associated with the decomposition (5.19) is

$$\boldsymbol{f}^{(1)} := V^{(1)} f\big(H^{(1)}[K^{(1)}]^{-1}\big)\|\boldsymbol{b}\|\boldsymbol{e}_1.$$

By Theorem 5.8 we know that $\boldsymbol{f}^{(1)} = r_m(A)\boldsymbol{b}$, where $r_m \in \mathcal{P}_{m-1}/q_{m-1}$ interpolates f at the nodes $\Lambda(H^{(1)}[K^{(1)}]^{-1})$.

It is now possible to restart the rational Arnoldi algorithm, thus generalizing the algorithm for restarted polynomial Krylov approximations presented by Eiermann & Ernst [EE06]. We let

$$AV^{(2)}K^{(2)} = V^{(2)}H^{(2)} + \boldsymbol{v}^{(2)}\boldsymbol{h}^{(2)}$$

be a reduced rational Krylov decomposition with starting vector $\boldsymbol{v}^{(1)}$, i.e., $V^{(2)}\boldsymbol{e}_1 = \boldsymbol{v}^{(1)}$. By appending this decomposition to the previous one (5.19) we obtain

$$A[V^{(1)}, V^{(2)}]\begin{bmatrix} K^{(1)} & \\ & K^{(2)} \end{bmatrix} = [V^{(1)}, V^{(2)}]\begin{bmatrix} H^{(1)} & \\ \boldsymbol{e}_1\boldsymbol{h}^{(1)} & H^{(2)} \end{bmatrix} + \boldsymbol{v}^{(2)}[\boldsymbol{0}^T, \boldsymbol{h}^{(2)}].$$

The rational Krylov approximation associated with this (again reduced) decomposition is

$$\boldsymbol{f}^{(2)} := [V^{(1)}, V^{(2)}] f\Big(\begin{bmatrix} H^{(1)} & \\ \boldsymbol{e}_1\boldsymbol{h}^{(1)} & H^{(2)} \end{bmatrix}\begin{bmatrix} K^{(1)} & \\ & K^{(2)} \end{bmatrix}^{-1}\Big)\|\boldsymbol{b}\|\boldsymbol{e}_1, \tag{5.20}$$

and Theorem 5.8 asserts that $\boldsymbol{f}^{(2)} = r_{2m}(A)\boldsymbol{b}$, where $r_{2m} \in \mathcal{P}_{2m-1}/q_{2m-1}$ interpolates f at the nodes $\Lambda(H^{(1)}[K^{(1)}]^{-1}) \cup \Lambda(H^{(2)}[K^{(2)}]^{-1})$.

Due to the block structure of the accumulated "H" and "K" matrices, the term $f(\cdots)$ in (5.20) again has a block structure, namely

$$\boldsymbol{f}^{(2)} = [V^{(1)}, V^{(2)}] \begin{bmatrix} f\big(H^{(1)}[K^{(1)}]^{-1}\big) & \\ F^{(2)} & f\big(H^{(2)}[K^{(2)}]^{-1}\big) \end{bmatrix} \|\boldsymbol{b}\| \boldsymbol{e}_1,$$

where $F^{(2)} \in \mathbb{C}^{m \times m}$. Hence we obtain an update formula

$$\begin{aligned} \boldsymbol{f}^{(2)} &= V^{(1)} f\big(H^{(1)}[K^{(1)}]^{-1}\big)\|\boldsymbol{b}\|\boldsymbol{e}_1 + V^{(2)}F^{(2)}\|\boldsymbol{b}\|\boldsymbol{e}_1 \\ &= \boldsymbol{f}^{(1)} + V^{(2)}F^{(2)}\|\boldsymbol{b}\|\boldsymbol{e}_1, \end{aligned}$$

which allows us to compute $\boldsymbol{f}^{(2)}$, a rational Krylov approximation of order $2m$, from $\boldsymbol{f}^{(1)}$ using only the last m Krylov basis vectors in $V^{(2)}$. This restarting procedure can be repeated until a sufficiently good rational Krylov approximation is obtained, never exceeding the storage requirement of m Krylov basis vectors (the number m is often called the *restart length*). A computational problem, however, is the evaluation of the matrix function $f(\cdots)$ for block matrices growing in size with each restart. This problem has been addressed in [AEEG08b] for the polynomial restart algorithm by replacing f with a suitable rational approximation of it. Other aspects related to restarted Krylov methods for approximating $f(A)\boldsymbol{b}$ are discussed in [AEEG08a, EEG09].

5.4.2 The PAIN Method

A particularly simple rational Krylov method is given by the iteration

$$\boldsymbol{v}_1 = \boldsymbol{b}/\|\boldsymbol{b}\|, \tag{5.21a}$$

$$\beta_j \boldsymbol{v}_{j+1} = (I - A/\xi_j)^{-1}(A - \alpha_j I)\boldsymbol{v}_j, \quad j = 1, \ldots, m, \tag{5.21b}$$

the numbers $\alpha_j, \beta_j \in \mathbb{C}$ being arbitrary for the moment, except that we require $\alpha_j \neq \xi_j$ and $\beta_j \neq 0$ for all $j = 1, \ldots, m$. It is easily seen that this iteration generates a rational

Krylov decomposition $AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m}$, where $V_{m+1} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m+1}]$ and

$$\underline{K_m} = \left[\begin{array}{ccccc} 1 & & & & \\ \beta_1/\xi_1 & 1 & & & \\ & \beta_2/\xi_2 & \ddots & & \\ & & \ddots & 1 & \\ \hline & & & & \beta_m/\xi_m \end{array}\right] \quad \text{and} \quad \underline{H_m} = \left[\begin{array}{ccccc} \alpha_1 & & & & \\ \beta_1 & \alpha_2 & & & \\ & \beta_2 & \ddots & & \\ & & \ddots & \alpha_m & \\ \hline & & & & \beta_m \end{array}\right].$$

For $\xi_m = \infty$ we obtain a reduced decomposition $AV_mK_m = V_{m+1}\underline{H_m}$. By Theorem 5.8 the associated rational Krylov approximation satisfies

$$\boldsymbol{f}_m^{\mathrm{RK}} = V_m f(H_m K_m^{-1}) \|\boldsymbol{b}\| \boldsymbol{e}_1 = r_m(A)\boldsymbol{b}, \tag{5.22}$$

where r_m is a rational function with poles $\xi_1, \ldots, \xi_{m-1}$ that interpolates f at the eigenvalues $\Lambda(H_m K_m^{-1})$, and these are obviously the points $\alpha_1, \ldots, \alpha_m$. Therefore the approximation method just described corresponds to rational interpolation of $f(A)\boldsymbol{b}$ with *preassigned poles and interpolation nodes*. To refer to this method easily we call it the *PAIN method*. Some comments are in order:

— The matrix $H_m K_m^{-1}$ is independent of the last pole ξ_m, and so is the approximation $\boldsymbol{f}_m^{\mathrm{RK}}$ in (5.22). Hence, for computing $\boldsymbol{f}_m^{\mathrm{RK}}$ it is not necessary to set $\xi_m = \infty$.

— The iteration (5.21) involves no inner products with Krylov basis vectors. The matrices H_m and K_m have the same bidiagonal structure as if they were computed by a restarted rational Arnoldi algorithm with restart length one. This means that only one Krylov basis vector needs to be stored at a time and the approximation $\boldsymbol{f}_m^{\mathrm{RK}}$ can be obtained via updating $\boldsymbol{f}_m^{\mathrm{RK}} = \boldsymbol{f}_{m-1}^{\mathrm{RK}} + c_m \boldsymbol{v}_m$, where c_m denotes the last entry of the vector $f(H_m K_m^{-1}) \|\boldsymbol{b}\| \boldsymbol{e}_1$.

— The PAIN method requires, in addition to the poles $\{\xi_j\}$, a suitable sequence of interpolation nodes $\{\alpha_j\}$. The choice of both sequences is discussed in Chapter 7.

— If all poles $\xi_j = \infty$, the PAIN method reduces to a polynomial interpolation method. If in addition all *scaling factors* $\beta_j = 1$, the PAIN method is equivalent to evaluating

an interpolation polynomial in Newton form

$$\boldsymbol{f}_m^{\mathrm{RK}} = \sum_{k=1}^{m} d_k \prod_{j=1}^{k-1} (A - \alpha_j I)\boldsymbol{b}. \tag{5.23}$$

Such a method is described in [HPKS99][4], where it is also proposed to choose the interpolation nodes α_j as Leja points in a compact set Σ containing $\Lambda(A)$. Therefore this method is also called the *Leja point method* [CVB04, BCV04, CVB07]. The Leja point method requires a stable computation of the divided differences d_k for f. Based on the observation that

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix} = f\Bigg(\begin{bmatrix} \alpha_1 & & & \\ 1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & 1 & \alpha_m \end{bmatrix} \Bigg) \boldsymbol{e}_1 \tag{5.24}$$

(see [Opi64, MNP84] or [Hig08, Thm. 10.22]), it is proposed in [Cal07] to determine the divided differences d_k by (5.24) via a matrix function evaluation and then to evaluate the polynomial (5.23). This is exactly what the polynomial PAIN method does implicitly (note that (5.24) equals $f(H_m K_m^{-1})\boldsymbol{e}_1$ if $\beta_j = 1$, $\xi_j = \infty$ for all $j = 1, \ldots, m$).

— To avoid overflow in Newton interpolation it is known that the set of interpolation points Σ needs to be scaled to have unit logarithmic capacity [Rei90, Tal91]. This rescaling is not required for the PAIN method if we choose the scaling factors β_j such that all Krylov vectors $\boldsymbol{v}_j$ have unit length.

Remark 5.9. The PAIN method can be combined with the Rayleigh–Ritz method to obtain a *hybrid rational Krylov algorithm* (in the style of [NRT92]) for problems of the form $f(A)\boldsymbol{b}^{(j)}$ with distinct vectors $\boldsymbol{b}^{(j)}$, $j = 0, 1, \ldots, p$. In the first phase of such a hybrid algorithm one computes a Rayleigh–Ritz approximation $\boldsymbol{f}_m^{(0)}$ for $f(A)\boldsymbol{b}^{(0)}$ and the associated rational Ritz values. In the second phase one runs in parallel p instances of the PAIN method for approximating $f(A)\boldsymbol{b}^{(j)}$, $j = 1, \ldots, p$, using as interpolation nodes the rational Ritz values from the first phase (suitably reordered).

[4] This article appeared in *The Journal of Chemical Physics* and seems to have been overlooked by the matrix function community. Another contribution of the authors is the proposal to insert A directly into a Faber expansion of f, an approach also advocated in [MN01b, Nov03, BCV03].

Remark 5.10. The PAIN method can easily be altered to approximate $f(A)$ instead of $f(A)\boldsymbol{b}$. To this end one formally sets $\boldsymbol{b} = I$ in (5.21) and replaces possible vector norms by operator norms. If $A \in \mathbb{C}^{N\times N}$ is a matrix, the "matricized" PAIN method is equivalent to approximating $f(A)\boldsymbol{e}_j$ for all unit coordinate vectors $\boldsymbol{e}_j \in \mathbb{C}^N$ simultaneously.

The observation that Krylov basis vectors can formally be replaced by operators naturally inspires the definition of *rational operator decompositions* of the form

$$A[A_1, \ldots, A_{m+1}]\underline{K_m} = [A_1, \ldots, A_{m+1}]\underline{H_m},$$

where $A, A_1, \ldots, A_{m+1}$ are bounded linear operators on a Banach or Hilbert space. It would be interesting to study such decompositions (e.g., what is the interpretation of the eigenvalues $\Lambda(H_m K_m^{-1})$ associated with these decompositions?), but this is beyond the scope of this thesis.

5.4.3 The Shift-and-Invert Method

The *shift-and-invert method* for the approximation of matrix functions was introduced independently by Moret & Novati [MN04] (there referred to as the *restricted denominator method*) and van den Eshof & Hochbruck [EH06]. The principle of this method is to run the polynomial Arnoldi algorithm with the spectrally transformed operator $(A - \xi I)^{-1}$, $\xi \notin \Lambda(A)$, which yields a polynomial Arnoldi decomposition

$$(A - \xi I)^{-1} V_m = V_{m+1}\underline{H_m}, \quad \text{where} \quad \underline{H_m} = \begin{bmatrix} H_m \\ h_{m+1,m}\boldsymbol{e}_m^T \end{bmatrix} \in \mathbb{C}^{(m+1)\times m} \tag{5.25}$$

is an unreduced upper Hessenberg matrix, and $V_{m+1} = [V_m, \boldsymbol{v}_{m+1}]$ is an orthonormal basis of $\mathcal{Q}_{m+1} = \mathcal{K}_{m+1}((A - \xi I)^{-1}, \boldsymbol{b})$, a rational Krylov space with all poles at ξ. The *shift-and-invert approximation* for $f(A)\boldsymbol{b}$ is defined as

$$\boldsymbol{f}_m^{\mathrm{SI}} := V_m f(S_m) V_m^* \boldsymbol{b} \quad \text{with} \quad S_m := H_m^{-1} + \xi I_m.$$

Note that (5.25) is equivalent to

$$AV_{m+1}\underline{H_m} = V_{m+1}(\xi\underline{H_m} + \underline{I_m}), \tag{5.26}$$

which is a special case of the rational Arnoldi decomposition (5.8) with $X_m = \xi I_m$ and $U_m = I_m$. Moreover, we have $S_m = (\xi H_m + I_m)H_m^{-1}$ so that the shift-and-invert approximation is a rational Krylov approximation (cf. (5.17)) associated with the decomposition (5.26). Unfortunately, this decomposition (5.26) is not reduced and hence Theorem 5.8 cannot be applied.

Note that S_m is not a Rayleigh quotient for (A, V_m): left-multiplying (5.25) by $V_m^*(A-\xi I)$ and separating the term $V_m^*AV_m$ yields

$$\begin{aligned} V_m^*AV_m &= H_m^{-1} + \xi I_m - V_m^*A\boldsymbol{v}_{m+1}h_{m+1,m}\boldsymbol{e}_m^TH_m^{-1} \\ &= S_m - V_m^*A\boldsymbol{v}_{m+1}h_{m+1,m}\boldsymbol{e}_m^TH_m^{-1} \\ &=: S_m - M_m, \end{aligned}$$

i.e., S_m is a rank-1 modification of the Rayleigh quotient $A_m = V_m^*AV_m$ with

$$M_m = V_m^*A\boldsymbol{v}_{m+1}h_{m+1,m}\boldsymbol{e}_m^TH_m^{-1}. \tag{5.27}$$

Therefore $\boldsymbol{f}_m^{\mathrm{SI}}$ is not a Rayleigh–Ritz approximation for $f(A)\boldsymbol{b}$ from $\mathcal{Q}_m$. However, with the function $\hat{f}(\hat{z}) := f(\hat{z}^{-1} + \xi)$ we have $\boldsymbol{f}_m^{\mathrm{SI}} = V_m\hat{f}(H_m)V_m^*\boldsymbol{b}$, hence $\boldsymbol{f}_m^{\mathrm{SI}}$ is a *polynomial* Rayleigh–Ritz approximation for $\hat{f}((A-\xi I)^{-1})\boldsymbol{b} = f(A)\boldsymbol{b}$ associated with the Arnoldi decomposition (5.25). This connection allows us to conclude from Lemma 3.10 that there exists an interpolation characterization of the shift-and-invert approximation (cf. [EH06]).

Theorem 5.11. *There holds $\boldsymbol{f}_m^{\mathrm{SI}} = r_m(A)\boldsymbol{b}$, where $r_m(z) = p_{m-1}(z)/(z-\xi)^{m-1}$ interpolates f at the nodes $\Lambda(S_m)$.*

A near-optimality result similar to Theorem 4.10 can also be given.

Theorem 5.12. *If f is analytic in a neighborhood of $\mathbb{W}(A)$, then for every set $\Sigma \supseteq \mathbb{W}(A)$ there holds*

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{\mathrm{SI}}\| \le 2C\|\boldsymbol{b}\| \min_{p_{m-1}\in\mathcal{P}_{m-1}} \|f(z) - p_{m-1}(z)/(z-\xi)^{m-1}\|_\Sigma,$$

with a constant $C \le 11.08$.

If A is self-adjoint the result holds with $C = 1$.

Let A be a self-adjoint operator, in which case the matrices A_m and $S_m = H_m^{-1} + \xi I_m$ are clearly Hermitian. Moreover, these matrices are nonderogatory and hence the real eigenvalues $\Lambda(A_m)$ and $\Lambda(S_m)$ must be distinct, respectively. Let us have a closer look at these eigenvalues, which are plotted in Figure 5.1 for a simple example. One observes an interlacing property in the sense that between any two neighboring rational Ritz values in $\Lambda(A_m)$ we find exactly one point of $\Lambda(S_m)$, and vice versa. Additionally, the following theorem asserts that the eigenvalues of S_m are always "closer" to the real pole ξ than the Ritz values.

Theorem 5.13. *Let A be self-adjoint with $\Lambda(A) \subseteq [a, b]$. Let $\theta_1 < \theta_2 < \cdots < \theta_m$ denote the rational Ritz values $\Lambda(A_m)$ associated with the rational Krylov space generated with a single repeated pole $\xi \in \mathbb{R} \setminus [a, b]$, and let $\sigma_1 < \sigma_2 < \cdots < \sigma_m$ denote the eigenvalues $\Lambda(S_m)$ (which are the interpolation nodes underlying the shift-and-invert method). Then*

$$\begin{aligned} &\sigma_1 \le \theta_1 \le \sigma_2 \le \theta_2 \le \cdots \le \sigma_m \le \theta_m && \text{if } \xi < a, \\ &\theta_1 \le \sigma_1 \le \theta_2 \le \sigma_2 \le \cdots \le \theta_m \le \sigma_m && \text{if } b < \xi. \end{aligned}$$

Proof. By (5.26) we have

$$AV_m = V_m(\xi I_m + H_m^{-1}) + (\xi \boldsymbol{v}_{m+1} - A\boldsymbol{v}_{m+1}) h_{m+1,m} \boldsymbol{e}_m^T H_m^{-1}. \tag{5.28}$$

Let us verify that $V_m^* A \boldsymbol{v}_{m+1}$ is an eigenvector of the matrix M_m defined in (5.27):

$$\begin{aligned} M_m(V_m^* A \boldsymbol{v}_{m+1}) &= (V_m^* A \boldsymbol{v}_{m+1} h_{m+1,m} \boldsymbol{e}_m^T H_m^{-1})[AV_m]^* \boldsymbol{v}_{m+1} \\ &= (V_m^* A \boldsymbol{v}_{m+1} h_{m+1,m} \boldsymbol{e}_m^T H_m^{-1})[H_m^{-1} \boldsymbol{e}_m h_{m+1,m} (\xi \boldsymbol{v}_{m+1} - A\boldsymbol{v}_{m+1})^* \boldsymbol{v}_{m+1}] \\ &= (V_m^* A \boldsymbol{v}_{m+1}) h_{m+1,m}^2 \boldsymbol{e}_m^T H_m^{-2} \boldsymbol{e}_m (\xi - \boldsymbol{v}_{m+1}^* A \boldsymbol{v}_{m+1}) \\ &=: (V_m^* A \boldsymbol{v}_{m+1}) \mu_m, \end{aligned}$$

where we have used (5.28) for the substitution in square brackets. The Rayleigh quotient $\boldsymbol{v}_{m+1}^* A \boldsymbol{v}_{m+1}$ is clearly contained in the interval $[a, b]$ and hence the eigenvalue μ_m is negative for $\xi < a$ and positive for $b < \xi$. Therefore the inertia[5] of M_m is $(0, 1, m-1)$ and $(1, 0, m-1)$, respectively. The remainder of the proof is a direct application of the *rank theorem* [Par98, Cor. 10.3.1]. □

[5] The inertia of a matrix is the triple of the numbers of its positive, negative and zero eigenvalues, see [Par98, p. 11].

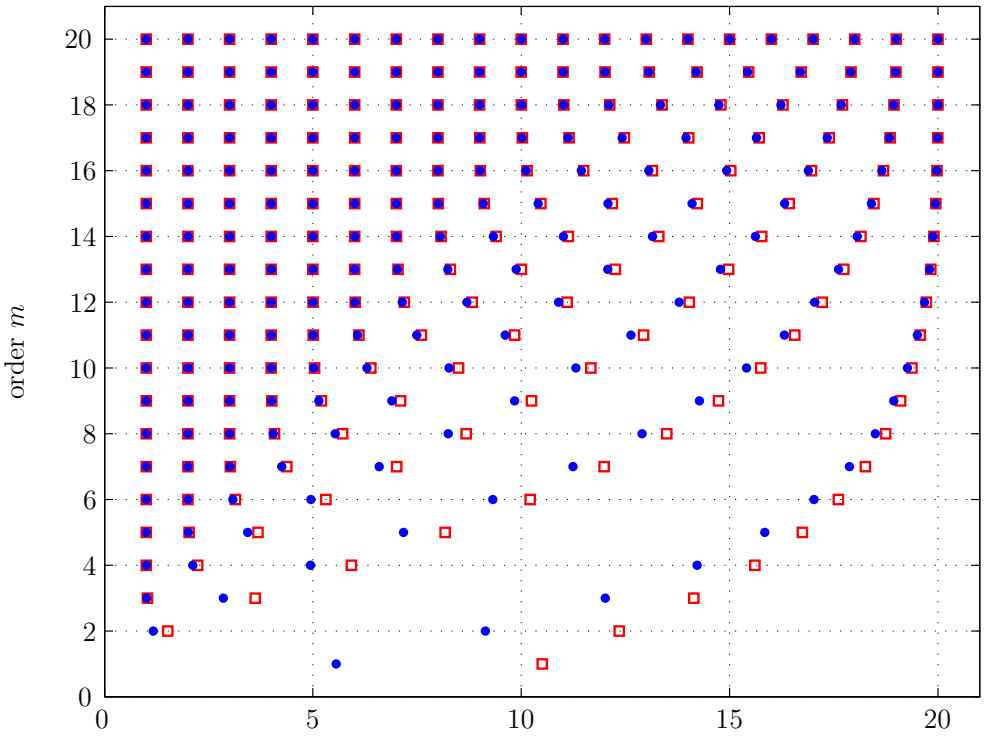


Figure 5.1: The red squares indicate the rational Ritz values $\Lambda(A_m)$, and the blue dots indicate the interpolation nodes $\Lambda(S_m)$ of the shift-and-invert method. In this example we have chosen $A = \mathrm{diag}(1, 2, \ldots, 20)$, $\boldsymbol{b} = [1, \ldots, 1]^T$, and $\xi = 0$.

5.4.4 The PFE Method

A convenient form of a rational function is its partial fraction expansion (PFE)

$$r_m(z) = \gamma_0 + \frac{\gamma_1}{z - \xi_1} + \cdots + \frac{\gamma_{m-1}}{z - \xi_{m-1}},$$

where the numbers γ_j are referred to as *residues*. Given such an expansion we can directly compute the vector

$$\boldsymbol{f}_m^{\mathrm{PFE}} := r_m(A)\boldsymbol{b} = \gamma_0 \boldsymbol{b} + \gamma_1 (A - \xi_1 I)^{-1} \boldsymbol{b} + \cdots + \gamma_{m-1} (A - \xi_{m-1} I)^{-1} \boldsymbol{b}, \tag{5.29}$$

an approach we will refer to as the *PFE method*. This method does not utilize a rational Krylov decomposition in the sense of Definition 5.5, but its "search space" is obviously a rational Krylov space $\mathcal{Q}_m(A, \boldsymbol{b})$ with poles $\xi_1, \ldots, \xi_{m-1}$. The quality of $\boldsymbol{f}_m^{\mathrm{PFE}}$ as an approximation for $f(A)\boldsymbol{b}$ depends on the rational function r_m, since by Crouzeix's theorem

(cf. Theorem 4.9) we have for every set $\Sigma \supseteq \mathbb{W}(A)$:

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{\mathrm{PFE}}\| \leq C\|\boldsymbol{b}\|\,\|f - r_m\|_\Sigma, \qquad C \leq 11.08.$$

Note that the $m-1$ shifted linear systems in (5.29) can be solved independently and hence the PFE method is perfectly suited for a parallel computer [GS89]. Efficient variants of the PFE method have appeared in the literature, e.g., using different polynomial Krylov methods such as CG [EFL$^+$02, FS08b] or restarted FOM [AEEG08b], or $\mathcal{H}$-matrix techniques [GHK02, GHK03, GHK04, GHK05], to solve these linear systems. The practical computation of near-best approximations r_m to f was greatly enhanced by Trefethen and coauthors [TWS06, ST07a, ST07b, HHT08] using the Carathéodory–Fejér method and contour integrals (cf. Chapter 7).

The applicability of the PFE method hinges on the ability to compute the rational function r_m and its partial fraction expansion in a stable way. This is in contrast to the Rayleigh–Ritz method, the PAIN method or any other rational Krylov method based on rational interpolation: errors in r_m are directly reflected in $\boldsymbol{f}_m^{\mathrm{PFE}}$ and there is no way to reduce these errors by iterating further. The PFE method is therefore not robust against perturbations. Or to say it in other words, any method based on rational interpolation of f is not very sensitive to perturbed poles and/or interpolation nodes, whereas a partial fraction is possibly very sensitive to inaccurate poles and residues. We remark that this problem has stimulated the development of techniques for computing *incomplete partial fraction expansions* [Hen71, Lau87, CGR95]. These representations are more stable but come at the price of losing parallelism.

Assume now we *can* stably compute a good partial fraction approximation r_m to f on Σ. Is there any reason to still use, e.g., Rayleigh–Ritz approximations?

The answer depends, of course, on the application. Here are a few reasons why the use of Rayleigh–Ritz approximations could be superior to the PFE method.

— In practice, most of the computation time is spent in solving shifted linear systems with A. Compared to this it often does not really matter whether one orthogonalizes m vectors or not. Once we have computed a rational Arnoldi decomposition we can use it to approximate $f(A)\boldsymbol{b}$ for whatever function we like.

— Rational Arnoldi decompositions, and more generally, rational Krylov decompositions contain valuable spectral information about A as a by-product.

— The poles ξ_j of the partial fraction r_m are dictated by the function f and the set Σ. If these poles are complex, the PFE method may introduce complex arithmetic into real problems (though there may exist remedies, e.g., if A is symmetric [AK00]). On the other hand, Rayleigh–Ritz approximations for $f(A)\boldsymbol{b}$ can achieve high accuracy even if the poles of the search space $\mathcal{Q}_m$ are chosen rather arbitrarily.

5.5 Overview of Rational Krylov Approximations

At this point, let us briefly review the various rational Krylov approximations we have studied so far (there won't be any more).

Rayleigh approximations $\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b}$ with the Rayleigh quotient $A_m = V_m^\dagger A V_m$ are defined for an arbitrary search space $\mathcal{V}_m$. These approximations are independent of the basis V_m.

Rayleigh–Ritz approximations $\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b}$ are Rayleigh approximations where the search space is a polynomial Krylov space $\mathcal{K}_m$ or, more generally, a rational Krylov space $\mathcal{Q}_m$. The underlying interpolation nodes are rational Ritz values $\Lambda(A_m)$.

Rational Krylov approximations $\boldsymbol{f}_m^{\mathrm{RK}} = V_m f(H_m K_m^{-1}) V_m^\dagger \boldsymbol{b}$ are associated with a rational Krylov decomposition (5.15) and their search space is a rational Krylov space $\mathcal{Q}_m$. These approximations are dependent on the basis V_m. In case that the rational Krylov decomposition is reduced (cf. (5.16)), the interpolation nodes underlying $\boldsymbol{f}_m^{\mathrm{RK}}$ are $\Lambda(H_m K_m^{-1})$, and if this reduced decomposition is orthonormal then $\boldsymbol{f}_m^{\mathrm{RK}}$ coincides with the Rayleigh–Ritz approximation $\boldsymbol{f}_m$ from $\mathcal{Q}_m$.

Shift-and-invert approximations $\boldsymbol{f}_m^{\mathrm{SI}}$ are special rational Krylov approximations extracted from rational Krylov spaces generated with a single repeated pole.

Partial fraction approximations $\boldsymbol{f}_m^{\mathrm{PFE}} = r_m(A)\boldsymbol{b}$ are computed by direct evaluation of a rational function r_m, which is (in some sense) a good approximation to f.

6 Computational Issues

Note: the minors are computed following this longer rule to avoid small differences of large numbers causing the loss of accuracy.
A. N. Krylov [Kry31]

We will address several questions related to the practical implementation of rational Krylov methods for approximating $f(A)\boldsymbol{b}$. We begin with some remarks on the efficient computation of Rayleigh quotients in Section 6.1. In Section 6.2 we discuss the important issue of solving linear systems in a rational Krylov method, followed by the derivation of an error estimate for Rayleigh–Ritz approximations computed with inexact linear system solves in Section 6.3. In Section 6.4 we have some remarks about the loss of orthogonality in the rational Arnoldi algorithm, and in Section 6.5 we investigate several parallel variants of this algorithm. Section 6.6 is devoted to a-posteriori error estimates for Rayleigh–Ritz approximations.

This chapter also includes a few numerical toy problems with the function $f(z) = \exp(z)$. The main purpose of these examples is to illustrate our findings, which are neither limited to small problems nor to a particular function f. More extensive numerical computations are the subject of Chapter 9.

6.1 Obtaining the Rayleigh Quotient

For the efficient computation of a Rayleigh–Ritz approximation $\boldsymbol{f}_m = V_m f(A_m) V_m^\dagger \boldsymbol{b}$ from a rational Krylov space $\mathcal{Q}_m$ it is crucial to obtain the Rayleigh quotient $A_m = V_m^\dagger A V_m$ cheaply. In what follows we discuss several approaches in this direction.

The Explicit Projection Approach. Let us assume that V_m is an ascending orthonormal basis as is computed, e.g., by the rational Arnoldi algorithm (cf. Algorithm 5.1 on page 40). In this case we have $A_m = V_m^* A V_m$ and the most straightforward approach for computing A_m is to use this projection formula explicitly. To make this projection as efficient as possible, one can exploit the fact that the Rayleigh quotient A_{m-1} is a leading principal submatrix of A_m and hence only the last column and row vectors of A_m need to be computed in iteration m. If A is self-adjoint then so is A_m and the computation of only one of these vectors is required.

We also recall that in iteration m of the rational Arnoldi algorithm the vector

$$\boldsymbol{x}_m = (I - A/\xi_m)^{-1} A \boldsymbol{y}_m, \quad \boldsymbol{y}_m = \sum_{i=1}^{m} \boldsymbol{v}_i u_{i,m} \tag{6.1}$$

is computed. If we store in addition to the Krylov basis V_m a matrix $Z_m = [\boldsymbol{z}_m, \ldots, \boldsymbol{z}_m] := AV_m$ we have

$$\boldsymbol{x}_m = (I - A/\xi_m)^{-1} \widetilde{\boldsymbol{y}}_m, \quad \widetilde{\boldsymbol{y}}_m = \sum_{i=1}^{m} \boldsymbol{z}_i u_{i,m},$$

whose evaluation requires the same number of arithmetic operations as (6.1). The matrix Z_m is useful because the last row of A_m is $\boldsymbol{v}_m^* Z_m$ and the last column is $V_m^* \boldsymbol{z}_m$, hence no additional products with A are required for computing A_m. To update Z_m to $Z_{m+1} = [Z_m, A\boldsymbol{v}_{m+1}]$ one only has to append the column $A\boldsymbol{v}_{m+1}$ upon availability of $\boldsymbol{v}_{m+1}$.

Performing a Polynomial Krylov Step. An efficient approach for computing the Rayleigh quotient A_m is proposed by Beckermann & Reichel [BR09]. By Lemma 5.6 we know that A_m can be computed from the reduced rational Arnoldi decomposition $AV_m K_m = V_{m+1} \underline{H_m}$ as $A_m = H_m K_m^{-1}$. Such a reduced decomposition is obtained from the rational Arnoldi algorithm by setting $\xi_m = \infty$ (since the last row of $\underline{K_m}$ in (5.6) vanishes), i.e., by performing a polynomial Krylov step in iteration m.

In practice one may want to compute the Rayleigh quotients A_j during many iterations $j \leq m$ without setting all the corresponding poles $\xi_j = \infty$. A possible remedy is to perform a polynomial Krylov step by computing an *intermediate basis vector* $\widetilde{\boldsymbol{x}}_j = A\boldsymbol{y}_j$ using the continuation vector $\boldsymbol{y}_j$ in iteration j. This vector $\widetilde{\boldsymbol{x}}_j$ is then orthonormalized against the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j\}$ to complete the reduced rational Arnoldi decomposition $AV_jK_j = V_{j+1}\underline{H_j}$ from which $A_j = H_jK_j^{-1}$ is easily computed. Afterwards the vector $\widetilde{\boldsymbol{x}}_j$ can be reused to continue the rational Arnoldi algorithm with

$$\boldsymbol{x}_j = (I - A/\xi_j)^{-1}A\boldsymbol{y}_j = (I - A/\xi_j)^{-1}\widetilde{\boldsymbol{x}}_j.$$

With this naive approach both vectors $\widetilde{\boldsymbol{x}}_j$ and $\boldsymbol{x}_j$ need to be orthonormalized against the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j\}$, which is of course not desirable. We therefore propose the use of an *auxiliary basis vector* $\boldsymbol{v}_\infty$, which is initialized as $\boldsymbol{v}_\infty = A\boldsymbol{v}_1$ in the first iteration of the rational Arnoldi algorithm and permanently kept orthonormal to the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j$ in all iterations $j > 1$ (which requires only one additional orthogonalization per iteration). The Rayleigh quotient $A_j = \widetilde{H}_j\widetilde{K}_j^{-1}$ is then obtained from the auxiliary reduced Arnoldi decomposition $AV_j\widetilde{K}_j = [V_j, \boldsymbol{v}_\infty]\underline{\widetilde{H}_j}$. Afterwards this decomposition can be overwritten by the usual rational Arnoldi algorithm.

To summarize, we give a possible implementation of a rational Arnoldi algorithm, which computes the Rayleigh quotient in each iteration by explicit projection (yielding $\widehat{A}_j$) *and* by using an auxiliary basis vector $\boldsymbol{v}_\infty$ (yielding $\widetilde{A}_j$). For brevity we have used MATLAB indexing in Algorithm 3, e.g., $U_{1:j,1:j-1}$ denotes the upper $j \times (j-1)$ part of U_m. Note that in exact arithmetic there holds $\widehat{A}_j = \widetilde{A}_j$. However, we will see in Section 6.3 that it proves useful to compute both Rayleigh quotients separately.

Algorithm 3: Rational Arnoldi with computation of Rayleigh quotients $\widehat{A}_j$ and $\widetilde{A}_j$.

Input: $A, \boldsymbol{b}, \{\xi_1, \ldots, \xi_m\}, U_m$

$\boldsymbol{v}_1 := \boldsymbol{b}/\|\boldsymbol{b}\|$

$\widehat{A}_0 := [\ ]$; // empty matrix

for $j = 1, \ldots, m$ **do**

  $\boldsymbol{z}_j := A\boldsymbol{v}_j$

  $\widehat{A}_j := \begin{bmatrix} \widehat{A}_{j-1} & V_{j-1}^* \boldsymbol{z}_j \\ \boldsymbol{v}_j^* Z_{j-1} & \boldsymbol{v}_j^* \boldsymbol{z}_j \end{bmatrix}$; // = RQ by explicit projection

  if $j = 1$ **then**

    $\boldsymbol{v}_\infty := \boldsymbol{z}_1$; // initialize $\boldsymbol{v}_\infty$

  $h_{j,\infty} := \langle \boldsymbol{v}_\infty, \boldsymbol{v}_j \rangle$; // keep $\boldsymbol{v}_\infty$ orthogonal to V_j

  $\boldsymbol{v}_\infty := \boldsymbol{v}_\infty - \boldsymbol{v}_j h_{j,\infty}$

  $\widetilde{H}_j := \begin{bmatrix} \underline{H_{j-1}}, h_{1:j,\infty} \end{bmatrix}$; // auxiliary $\widetilde{H}_j$ and $\widetilde{K}_j$ such

  $\widetilde{K}_j := [U_{1:j,1:j-1}, \boldsymbol{e}_1] + \widetilde{H}_j \operatorname{diag}(\xi_1^{-1}, \ldots, \xi_{j-1}^{-1}, 0)$; // that $AV_j\widetilde{K}_j = [V_j, \boldsymbol{v}_\infty]\underline{\widetilde{H}_j}$

  $\widetilde{A}_j := \widetilde{H}_j \widetilde{K}_j^{-1}$; // = RQ from the decomposition

  $\widetilde{\boldsymbol{y}} := \sum_{i=1}^{j} \boldsymbol{z}_i u_{i,j}$; // usual rational Arnoldi

  $\boldsymbol{x} := (I - A/\xi_j)^{-1}\widetilde{\boldsymbol{y}}$

  for $i = 1, \ldots, j$ **do**

    $h_{i,j} := \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle$

    $\boldsymbol{x} := \boldsymbol{x} - \boldsymbol{v}_i h_{i,j}$

  $h_{j+1,j} := \|\boldsymbol{x}\|$

  $\boldsymbol{v}_{j+1} := \boldsymbol{x}/h_{j+1,j}$

Recursive Updating. Assume that the Rayleigh quotient A_m is known and we have a (not necessarily reduced) rational Arnoldi decomposition

$$AV_{m+1} \begin{bmatrix} K_m \\ \boldsymbol{k}_m^T \end{bmatrix} = V_{m+1} \begin{bmatrix} H_m \\ \boldsymbol{h}_m^T \end{bmatrix},$$

where $V_{m+1} = [V_m, \boldsymbol{v}_{m+1}]$ is orthonormal, K_m and H_m are $m \times m$ matrices, and $\boldsymbol{k}_m^T, \boldsymbol{h}_m^T \in \mathbb{C}^{1\times m}$. Multiplying this decomposition on the left by V_{m+1}^* we find

$$\begin{bmatrix} A_m & \boldsymbol{a}_1 \\ \boldsymbol{a}_2^T & a_{m+1} \end{bmatrix} \begin{bmatrix} K_m \\ \boldsymbol{k}_m^T \end{bmatrix} = \begin{bmatrix} H_m \\ \boldsymbol{h}_m^T \end{bmatrix},$$

which are two equations for the missing column $\boldsymbol{a}_1$ and row $\boldsymbol{a}_2^T$ of the Rayleigh quotient A_{m+1}, provided that the scalar a_{m+1} is known, which comes at the cost of one additional inner product in the rational Arnoldi algorithm. If A is self-adjoint then so is A_{m+1} and additional savings are possible.

We remark that this approach is possibly unstable if the Rayleigh quotient is updated for many iterations. The recursion is useful if one needs to extrapolate Rayleigh quotients A_{m+j} from A_m for a few iterations $j = 1, 2, \ldots$ only.

6.2 Linear System Solvers

At the core of a rational Krylov algorithm are linear systems of the form

$$(A - \xi I)\boldsymbol{x} = \boldsymbol{y}, \quad \text{or more generally,} \quad (A - \xi M)\boldsymbol{x} = \boldsymbol{y}, \tag{6.2}$$

where A and M are typically large sparse matrices having similar nonzero patterns. The efficient solution of these linear systems is crucial for a good overall performance of any rational Krylov method for approximating $f(A)\boldsymbol{b}$.

The solution of a linear system is probably the most important task in scientific computing. It is clearly beyond the scope of this thesis to discuss all possible approaches that may be applicable for solving shifted linear systems of the form (6.2). Instead we briefly discuss different types of solvers, emphasizing possible savings that arise from the special structure of (6.2). Existing methods of solution can be categorized into *direct* methods and *iterative* (or *relaxation*) methods.

6.2.1 Direct Methods

The computational kernel of direct methods is Gaussian elimination, i.e., the LU factorization of the system matrix $A - \xi I$ or $A - \xi M$, respectively. The operation of a sparse direct solver can be roughly divided into four steps [DDSV98, Ch. 6], namely

— a *reordering step* that permutes the rows and columns such that the LU factors suffer little fill, or that the matrix has special structure, such as block-triangular form,

— an *analysis step* (sometimes also referred to as *symbolic factorization*) that determines the nonzero structures of the LU factors and creates suitable data structures for them,

— the *numerical factorization* that actually computes the LU factorization,

— the *solve step* that performs forward and back substitution using the LU factors.

The first three steps are independent of the right-hand side $\boldsymbol{y}$ of the linear system (6.2). Therefore direct solvers are particularly effective if the same system needs to be solved for many right-hand sides $\boldsymbol{y} = \boldsymbol{y}_j$, which is the case in a rational Krylov method if the poles $\xi = \xi_j$ do not vary often. The first two steps depend on the sparsity structure of the system matrix only and hence need be done exactly once even if the pole $\xi = \xi_j$ changes.

If sparse direct solvers are applicable it is certainly a good idea to give them a try because tremendous progress has been made in this direction in recent decades. Nowadays highly parallel codes tailored for different matrix properties (symmetry, diagonal dominance, block structure, etc.) and computer architectures are available. For a comprehensive overview of sparse direct solution techniques we refer to [Duf97]. More recent benchmarks of modern direct solvers are reported in [Gup02, GSH07]. In our numerical experiments in Chapter 9 we use the code PARDISO [SG04, SG06] which, according to these benchmarks, seems to offer a good overall performance as a black-box solver considering computation time, parallelism, and memory requirements. The main drawbacks of direct solvers compared to iterative solvers are certainly the need for an explicit matrix representation of the operator A and larger memory requirements compared to iterative methods.

6.2.2 Iterative Methods

Geometric Multigrid. Multigrid methods for elliptic boundary value problems are iterative methods whose guiding principle is to smooth out error components in the solution vector which are, on the actual grid, highly oscillatory in the eigenvector basis of the system matrix A. Since a shift $A - \xi I$ does not affect the eigenvectors of A, we may expect that the same smoothing principle should apply for shifted linear systems. Let us illustrate that indeed the same smoother, here the damped Jacobi iteration, can be applied almost independently of the shift ξ for the 1D model problem (see, e.g., [Bri87])

$$x''(t) - \xi x(t) = y(t) \qquad \text{for } t \in (0,1), \tag{6.3a}$$

$$x(0) = x(1) = 0. \tag{6.3b}$$

The finite-difference discretization of (6.3) in the points $\Omega_n := \{j/(n+1) : j = 1, \ldots, n\}$ yields a linear system $(A - \xi I)\boldsymbol{x} = \boldsymbol{y}$, where

$$A = (n+1)^2 \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad \boldsymbol{x},\, \boldsymbol{y} \in \mathbb{R}^n.$$

Let us split the system matrix $A - \xi I = D - R$ into the diagonal $D = (2(n+1)^2 - \xi)I$ and the remaining matrix R. The damped Jacobi iteration for $(A - \xi I)\boldsymbol{x} = \boldsymbol{y}$ is given as $\boldsymbol{x}_{m+1} = P_\omega \boldsymbol{x}_m + \omega D^{-1}\boldsymbol{y}$, where $0 < \omega \leq 1$ is the *damping parameter* and

$$P_\omega = (1-\omega)I + \omega D^{-1}R = (1-\omega)I + \omega \frac{(n+1)^2}{2(n+1)^2 - \xi} \begin{bmatrix} 0 & -1 & & \\ -1 & 0 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 0 \end{bmatrix}$$

is the *Jacobi iteration matrix*, the eigenvalues of which are

$$\Lambda(P_\omega) = \left\{ \lambda_j = 1 - \omega + \omega \frac{(n+1)^2}{(n+1)^2 - \xi} \cos\left(\frac{\pi j}{n+1}\right) : j = 1, \ldots, n \right\}.$$

It is known that the absolute values of the eigenvalues $\Lambda(P_\omega)$ determine the rate of error reduction in the direction of the associated eigenvectors (which are the same as the eigenvectors of $A - \xi I$). Note that the eigenvalues $\Lambda(P_\omega)$ are only slightly affected by the shift ξ if n is sufficiently large. Hence we expect that the Jacobi iteration on fine grids will converge just as if the shift ξ were absent. In particular, the optimal damping parameter $\omega = 2/3$ known for the unshifted problem (6.3) with $\xi = 0$ should also be a reasonable choice for the shifted problem.

Apart from the smoother, the other important components of a multigrid method are so-called *intergrid transfers*, which are matrices that map vectors associated with the *coarse grid* Ω_n to vectors associated with the *fine grid* Ω_{2n} and vice versa (*interpolation* and *restriction*). Since the grids depend mainly on the geometry of the problem and its discretization, the same intergrid transfers can be used for various shifted systems $A - \xi_j I$.

Polynomial Krylov Methods. The shift-invariance $\mathcal{K}_m(A, \boldsymbol{b}) = \mathcal{K}_m(A - \xi I, \boldsymbol{b})$ of polynomial Krylov spaces has led to the development of shifted-system versions for virtually every polynomial Krylov solver [Fre90, Fre93, FG98, Fro03, Sim03]. Unfortunately, these methods are not useful as linear system solvers within a rational Krylov method, at least in unpreconditioned form: solving the linear systems of a rational Krylov method with an unpreconditioned polynomial Krylov method is equivalent to working in a subspace of a polynomial Krylov space. As a simple example, assume we have computed a basis V_m of a "rational Krylov space" $\mathcal{Q}_m$ using s iterations of a polynomial Krylov method to solve each of the shifted linear systems associated with the basis vectors. All together, we have thus required $(m-1)s$ operator-vector products (the vector $\boldsymbol{b} \in \mathcal{Q}_m$ requires no linear system solve) and therefore $\mathcal{Q}_m \subset \mathcal{K}_{(m-1)s}$. Note that $\mathcal{Q}_m$ is *not* a rational Krylov space because the linear systems were solved inexactly. By Theorem 4.10 we know that the Rayleigh–Ritz approximation for $f(A)\boldsymbol{b}$ from $\mathcal{K}_{(m-1)s}$ is near-optimal, hence we expect that any extraction from the subspace $\mathcal{Q}_m \subset \mathcal{K}_{(m-1)s}$ is worse.

Obviously, polynomial linear system solvers need preconditioning in order to be efficient within a rational Krylov method. If sequences of (shifted) linear systems are to be solved it is desirable to have preconditioners that can be updated cheaply from one linear system to the next, taking advantage of previous computations. The design of such updating techniques is still an active area of research, see, e.g., the recent developments of updated approximate inverse preconditioners [BB03, Ber04, TT07].

Hierarchical Matrices. The idea of the hierarchical matrix technique is to compute a data-sparse approximation of a (generally dense) matrix A. *Data-sparse* means that only few data are needed for representing the approximation [HGB02]. This representation then allows one to compute matrix-vector products or to solve linear systems with almost linear complexity in the size of A, which consequently allows for the efficient use of polynomial or rational Krylov methods. Hierarchical matrix techniques have been applied successfully to the computation of the matrix exponential [GHK02], the matrix sign function [GHK03], and other matrix functions [GHK04, GHK05].

6.3 Inexact Solves

Many iterative methods are *inner-outer iterations* (also referred to as *two-stage methods* [FS92]), where the inner iteration is performed in an inexact way. Some examples are inexact Newton methods [DES82], inexact rational Krylov methods for eigenvalue problems [LLL97, LM98, SP99, GY00], and polynomial Krylov methods for the solution of linear systems of equations where the action of A is inexact [SS03, ES04, ESG05]. The theory of inexact Krylov methods has gained more attention in recent years. However, we are not aware of corresponding results for operator functions, except for some experiments reported in [EH06].

In the rational Arnoldi method, more precisely in (5.1), a linear system is solved. This is typically done approximately by an iterative method like multigrid or a preconditioned polynomial Krylov method, and hence we actually need to replace (5.1) by

$$\widetilde{\boldsymbol{x}}_j + \boldsymbol{d}_j = (I - A/\xi_j)^{-1} A \boldsymbol{y}_j, \quad \boldsymbol{y}_j = \sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j}, \tag{6.4}$$

where $\boldsymbol{d}_j := \boldsymbol{x}_j - \widetilde{\boldsymbol{x}}_j$ denotes the *error* of the approximate solution $\widetilde{\boldsymbol{x}}_j$. We assume here that the Gram–Schmidt orthogonalization is exact so that we only need to modify (5.2) to

$$\widetilde{\boldsymbol{x}}_j = \sum_{i=1}^{j+1} \boldsymbol{v}_i h_{i,j}. \tag{6.5}$$

The *residual* $\boldsymbol{r}_j := A\boldsymbol{y}_j - (I - A/\xi_j)\widetilde{\boldsymbol{x}}_j$ associated with (6.4) satisfies $(I - A/\xi_j)\boldsymbol{d}_j = \boldsymbol{r}_j$. By equating (6.4) and (6.5) and left-multiplying the result by $I - A/\xi_j$, we obtain a

decomposition

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m} + R_m,$$

where $\underline{H_m}$ and $\underline{K_m}$ are defined as in (5.6), $V_{m+1} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m+1}]$ has orthonormal columns, and $R_m = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m]$ is the *residual matrix*. Following [LM98] we rewrite this decomposition as

$$(A + D_m)V_{m+1}\underline{K_m} = V_{m+1}\underline{H_m}, \quad \text{where } D_m = -R_m\underline{K_m^\dagger}V_{m+1}^*. \tag{6.6}$$

We assert that this is a rational Arnoldi decomposition for the perturbed operator $A+D_m$, provided no pole ξ_j is contained in the spectrum $\Lambda(A + D_m)$. This assertion is verified by running the rational Arnoldi algorithm with the data $(A + D_m, \boldsymbol{b}, \{\xi_1, \ldots, \xi_m\})$ and showing that the approximate solutions $\widetilde{\boldsymbol{x}}_j$ from (6.4) satisfy exactly

$$\widetilde{\boldsymbol{x}}_j = \left(I - \frac{A + D_m}{\xi_j}\right)^{-1} (A + D_m)\boldsymbol{y}_j.$$

Indeed, the residual associated with this linear system is

$$\begin{aligned}
(A + D_m)\boldsymbol{y}_j - \left(I - \frac{A + D_m}{\xi_j}\right)\widetilde{\boldsymbol{x}}_j &= A\boldsymbol{y}_j - (I - A/\xi_j)\widetilde{\boldsymbol{x}}_j + D_m(\boldsymbol{y}_j + \widetilde{\boldsymbol{x}}_j/\xi_j) \\
&= \boldsymbol{r}_j - R_m\underline{K_m^\dagger}V_{m+1}^*(\boldsymbol{y}_j + \widetilde{\boldsymbol{x}}_j/\xi_j) \\
&= \boldsymbol{r}_j - R_m\underline{K_m^\dagger}(\underline{U_m}\boldsymbol{e}_j + \underline{H_m}\boldsymbol{e}_j/\xi_j) \\
&= \boldsymbol{r}_j - R_m\underline{K_m^\dagger}\underline{K_m}\boldsymbol{e}_j \\
&= \boldsymbol{0},
\end{aligned}$$

where we have used for the third equality the fact that the continuation vector $\boldsymbol{y}_j$ satisfies $\boldsymbol{y}_j = V_{m+1}\underline{U_m}\boldsymbol{e}_j$ with the upper triangular matrix $\underline{U_m} = [u_{i,j}] \in \mathbb{C}^{(m+1)\times m}$ by (6.4), and $\widetilde{\boldsymbol{x}}_j = V_{m+1}\underline{H_m}\boldsymbol{e}_j$ by (6.5).

Remark 6.1. It may indeed happen that poles ξ_j lie in the spectrum $\Lambda(A + D_m)$. As a simple example we consider the data

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \xi_1 = -1.$$

Clearly, $\boldsymbol{v}_1 = \boldsymbol{e}_1$ and if the linear system for $\boldsymbol{x}_1 = (I - A/\xi_1)^{-1}A\boldsymbol{v}_1$ is solved exactly, we

obtain after orthonormalization $\boldsymbol{v}_2 = [0, \sqrt{0.5}, -\sqrt{0.5}]^T$. However, if this linear system is solved with a residual $\boldsymbol{r}_1 = R_1 = [-1, -0.1, 0.1]^T$ we obtain a decomposition

$$(A + D_1)V_2\underline{K_1} = V_2\underline{H_1} + R_1 \quad \text{with } D_1 = \begin{bmatrix} -5 & 0 & 5 \\ -0.5 & 0 & 0.5 \\ 0.5 & 0 & -0.5 \end{bmatrix}$$

and it can easily be calculated that $A + D_1$ has an eigenvalue at -1. Hence, a rational Krylov space for the data $(A + D_1, \boldsymbol{b}, \{\xi_1\})$ does not exist. ——

We see from (6.6) that the Rayleigh quotient $\widetilde{A}_m$ computed with the (inexact) matrices $\underline{K_m}$ and $\underline{H_m}$ (cf. Section 6.1 for methods for doing this) is actually associated with the operator $A + D_m$, and not with A. Thus the approximation associated with (6.6),

$$\widetilde{\boldsymbol{f}}_m = V_m f(\widetilde{A}_m)V_m^*\boldsymbol{b} = V_m f(V_m^*(A + D_m)V_m)V_m^*\boldsymbol{b}, \tag{6.7}$$

is the Rayleigh–Ritz approximation for $f(A+D_m)\boldsymbol{b}$ from the rational Krylov space $\mathcal{Q}_m(A+D_m, \boldsymbol{b})$. We also refer to $\widetilde{\boldsymbol{f}}_m$ as an *inexact Rayleigh–Ritz approximation for* $f(A)\boldsymbol{b}$. The error $\|f(A)\boldsymbol{b} - \widetilde{\boldsymbol{f}}_m\|$ can be decomposed into the *sensitivity error* and the *approximation error* as

$$\|f(A)\boldsymbol{b} - \widetilde{\boldsymbol{f}}_m\| \le \underbrace{\|f(A)\boldsymbol{b} - f(A + D_m)\boldsymbol{b}\|}_{\text{sensitivity error}} + \underbrace{\|f(A + D_m)\boldsymbol{b} - \widetilde{\boldsymbol{f}}_m\|}_{\text{approximation error}}.$$

The sensitivity error may be estimated by sensitivity analysis of the function f, or by the estimator we will introduce soon, and the approximation error can be treated by error estimates for Rayleigh–Ritz approximations (cf. Section 6.6).

Example 6.2. To estimate the sensitivity error for the exponential function $f(z) = e^z$ one can use the formula (cf. [Bel70, p. 175], [Van77])

$$e^{A+D_m}\boldsymbol{b} - e^A\boldsymbol{b} = D_m \int_0^1 e^{(1-s)A}e^{s(A+D_m)}\boldsymbol{b}\,\mathrm{d}s,$$

to obtain

$$\|e^{A+D_m}\boldsymbol{b} - e^A\boldsymbol{b}\| \le \|D_m\| \cdot \|e^A\| \cdot \|e^{A+D_m}\boldsymbol{b}\| \lessapprox \|D_m\| \cdot \|e^A\|^2 \cdot \|\boldsymbol{b}\|.$$

——

By (6.6) we have

$$\|D_m\| \le \|R_m\| \cdot \|\underline{K_m}^{\dagger}\| \le \|R_m\|\, \sigma_{\min}^{-1}(\underline{K_m}),$$

where $\sigma_{\min}(\underline{K_m})$ denotes the smallest nonzero singular value of $\underline{K_m}$. Assume that all residuals satisfy $\|\boldsymbol{r}_j\| \le \tau$, then

$$\|R_m\| = \max_{\substack{\boldsymbol{y}\in\mathbb{C}^m \\ \|\boldsymbol{y}\|=1}} \Big\| \sum_{j=1}^{m} \boldsymbol{r}_j y_j \Big\| \le \max_{\substack{\boldsymbol{y}\in\mathbb{C}^m \\ \|\boldsymbol{y}\|=1}} \sum_{j=1}^{m} \|\boldsymbol{r}_j\| \cdot |y_j| = \tau \max_{\substack{\boldsymbol{y}\in\mathbb{C}^m \\ \|\boldsymbol{y}\|=1}} \|\boldsymbol{y}\|_1 \le \tau\sqrt{m}.$$

This inequality is sharp if and only if the residuals $\boldsymbol{r}_j$ are all collinear. In practice, this bound is often too crude and the norms $\|R_m\|$ stay of modest size for all orders m. Hence the norm $\|D_m\|$ and the sensitivity error are mainly determined by $\sigma_{\min}(\underline{K_m})$. Thus the rational Arnoldi algorithm can be considered backward stable if $\sigma_{\min}(\underline{K_m})$ does not become too small, a condition that can be easily monitored during the iteration.

A very small value of $\sigma_{\min}(\underline{K_m})$ indicates that one has computed a rational Arnoldi decomposition for the operator $A + D_m$ that is possibly "very far" from A, with the consequence that $\widetilde{\boldsymbol{f}}_m$ in (6.7) is possibly a very inaccurate approximation for $f(A)\boldsymbol{b}$.

An Estimator for the Sensitivity Error. In practice it is often observed that the Rayleigh quotient $\widehat{A}_m = V_m^* A V_m$ obtained by explicit projection with the basis V_m yields a *corrected approximation*

$$\widehat{\boldsymbol{f}}_m = V_m f(\widehat{A}_m) V_m^* \boldsymbol{b},$$

which is slightly closer to $f(A)\boldsymbol{b}$ than the approximation $\widetilde{\boldsymbol{f}}_m$ in (6.7). In general, $\widehat{\boldsymbol{f}}_m$ is *not* a Rayleigh–Ritz approximation and therefore not representable as a rational function $r_m(A)\boldsymbol{b}$, $r_m \in \mathcal{P}_{m-1}/q_{m-1}$. However, $\widehat{\boldsymbol{f}}_m$ is a Rayleigh approximation for $f(A)\boldsymbol{b}$ whereas $\widetilde{\boldsymbol{f}}_m$ is a Rayleigh–Ritz approximation for the perturbed problem $f(A+D_m)\boldsymbol{b}$. This observation inspires us to use $\|\widehat{\boldsymbol{f}}_m - \widetilde{\boldsymbol{f}}_m\|$ as an estimate for the sensitivity error, i.e.,

$$\|f(A)\boldsymbol{b} - f(A + D_m)\boldsymbol{b}\| \approx \|\widehat{\boldsymbol{f}}_m - \widetilde{\boldsymbol{f}}_m\| = \|f(\widehat{A}_m)V_m^*\boldsymbol{b} - f(\widetilde{A}_m)V_m^*\boldsymbol{b}\|. \tag{6.8}$$

It is easy to evaluate this estimate if the Rayleigh quotients $\widehat{A}_m$ and $\widetilde{A}_m$ are available, e.g., by the implementation of the rational Arnoldi algorithm in Section 6.1, page 64. In practice, one usually observes that (6.8) is a rapidly increasing curve that stagnates at

the level of the sensitivity error. Of course, when the approximation error falls below the sensitivity error (say, in iteration m_0) it is advised to stop the iteration because one would only improve approximations to a sequence of "wrong" problems $\{f(A+D_m)\boldsymbol{b}\}_{m>m_0}$.

Sometimes the computation of $\widehat{A}_m$ by explicit projection is not feasible, e.g., if A originates from a finite-element discretization and itself involves a linear system $A = M^{-1}K$. In this case one can make use of the residuals R_m since

$$\begin{aligned}\widehat{A}_m &= V_m^* A V_m \\ &= V_m^*(A+D_m)V_m - V_m^* D_m V_m \\ &= \widetilde{A}_m + V_m^*(R_m \underline{K_m^\dagger} V_{m+1}^*)V_m \\ &= \widetilde{A}_m + V_m^* R_m \underline{K_m^\dagger I_m},\end{aligned}$$

where $\underline{I_m}$ denotes the identity matrix of order m with an appended column of zeros. Thus, in the mth iteration of the rational Arnoldi algorithm we "only" need to perform $2m-1$ additional inner products

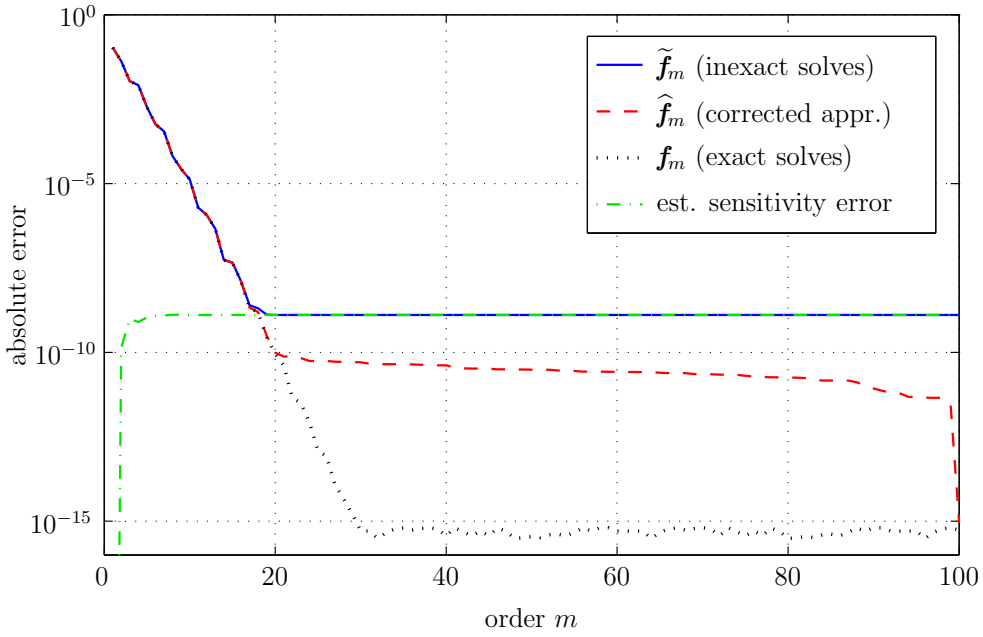
$$\boldsymbol{v}_1^*\boldsymbol{r}_m, \boldsymbol{v}_2^*\boldsymbol{r}_m, \ldots, \boldsymbol{v}_m^*\boldsymbol{r}_m \quad \text{and} \quad \boldsymbol{v}_m^*\boldsymbol{r}_1, \boldsymbol{v}_m^*\boldsymbol{r}_2, \ldots, \boldsymbol{v}_m^*\boldsymbol{r}_{m-1}$$

to obtain the corrected Rayleigh quotient $\widehat{A}_m$ without explicit projection (of course we exploit the fact that $V_{m-1}^* R_{m-1}$ is a leading principal submatrix of $V_m^* R_m$).

Example 6.3. To illustrate our sensitivity error estimate (6.8) we consider the computation of $f(A)\boldsymbol{b}$, $f(z) = \exp(z)$, with the simple data

$$A = \operatorname{diag}(-99, -98, \ldots, 0), \qquad \boldsymbol{b} = [1, \ldots, 1]^T, \qquad \xi_j = j \quad \text{for } j = 1, 2, \ldots \tag{6.9}$$

The error curves in Figure 6.1 show that the approximations $\widetilde{\boldsymbol{f}}_m$ (blue curve) obtained with inexact linear system solves in the rational Arnoldi algorithm (with residual norm 10^{-8}) stagnate at a higher level than if the linear systems were solved exactly (black dotted curve). We also observe that the corrected approximations $\widehat{\boldsymbol{f}}_m$ (red dashed curve) are better approximations for $f(A)\boldsymbol{b}$; in particular, $\widehat{\boldsymbol{f}}_{100}$ is exact to working precision. The sensitivity error estimate (6.8) (green dash-dotted curve) predicts well the stagnation level of the error curve $\|f(A)\boldsymbol{b} - \widetilde{\boldsymbol{f}}_m\|$ even for moderate orders m.



6inexact

Figure 6.1: Approximations for $f(A)\boldsymbol{b}$, $f(z) = \exp(z)$, with data (6.9). The linear systems in the rational Arnoldi algorithm were solved to a residual norm of 10^{-8}. The green dash-dotted curve shows the estimated sensitivity error (6.8).

6.4 Loss of Orthogonality

The results in the previous section suggest that losing orthogonality in V_{m+1} is somehow incompatible with inexact solves: our estimate (6.8) for the sensitivity error strongly relies on the orthogonality of V_{m+1}. In other words, we are not able to estimate the influence of inexact solves in the rational Arnoldi algorithm if we give up the orthogonality of V_{m+1}. On the other hand, if all linear systems are solved exactly and we have computed a (not necessarily orthogonal) reduced rational Krylov decomposition

$$AV_mK_m = V_{m+1}\underline{H_m},$$

then Theorem 5.8 asserts that the rational Krylov approximation $\boldsymbol{f}_m^{\mathrm{RK}} = V_m f(H_m K_m^{-1}) V_m^\dagger \boldsymbol{b}$ can be represented as a rational function interpolating f at the eigenvalues $\Lambda(H_m K_m^{-1})$. This characterization of $\boldsymbol{f}_m^{\mathrm{RK}}$ allows for an estimate of the approximation error $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{\mathrm{RK}}\|$, e.g., by making use of Crouzeix's theorem (cf. Theorem 4.9). If V_{m+1} is (nearly) orthonormal then the points $\Lambda(H_m K_m^{-1})$ are (nearly) rational Ritz values and we expect that the near-optimality result Theorem 4.10 is still applicable for practical purposes.

To sum up, we can either allow for inexact solves if we have orthogonality in V_{m+1}, or we can allow for loss of orthogonality in V_{m+1} if we guarantee that all linear systems in the rational Arnoldi algorithm are solved exactly. Unfortunately, we cannot allow for both, at least with the tools at our disposal. Fortunately, preventing loss of orthogonality in the rational Arnoldi decomposition is usually a minor problem because the Rayleigh–Ritz approximations $\boldsymbol{f}_m$ converge sufficiently fast that the number m of required Krylov basis vectors stays small and full reorthogonalization is still feasible. In our numerical experiments with the rational Arnoldi algorithm we found it sufficient to orthogonalize the Krylov basis vectors twice by the modified Gram–Schmidt algorithm, which is in agreement with W. Kahan's "twice is enough" rule (cf. [Par98, p. 115], [GLR02]).

6.5 Parallelizing the Rational Arnoldi Algorithm

The solution of large shifted linear systems is the bottleneck in a rational Krylov method, and hence it is advisable to parallelize it. Various possible implementations of the parallel rational Arnoldi algorithm were also discussed by Skoogh [Sko96, Sko98].

Given p processors, we mainly have two options for parallelizing the rational Arnoldi algorithm (cf. Algorithm 1 on page 42), namely

— *parallelism through partial fractions:* assign d distinct linear systems $(I - A/\xi_j)\boldsymbol{x}_j = A\boldsymbol{y}_j$ to different processors and solve them simultaneously ($d \leq p$),

— *parallelism at the solver level:* solve a single system $(I - A/\xi)\boldsymbol{x} = A\boldsymbol{y}$ by a parallel linear system solver on s processors ($s \leq p$).

These levels of parallelism are often referred to as *large grain parallelism* and *medium grain parallelism*, respectively [CGR95]. The so-called *low grain parallelism* is achieved by parallelizing the elementary arithmetic operations like, e.g., vector-vector sums or inner products. Although low grain parallelism is surely an important issue, we will only discuss medium and large grain parallelism here, referring to [KGGK94] for parallel treatment of low grain operations.

We will assume that $p = ds$ so that all processors are potentially busy with linear system solves. Note that the parallelism at the partial fraction level comes theoretically with

perfect speedup d compared to solving the linear systems sequentially on s processors, since the solution of these d linear systems is decoupled. Therefore it seems natural to maximize partial fraction parallelism, i.e., to make d as large as possible. An extreme case would be to have $d = p$ pairwise distinct poles $\xi_1, \ldots, \xi_d$ and to compute the vectors

$$\boldsymbol{x}_1 = (I - A/\xi_1)^{-1} A\boldsymbol{v}_1, \ldots, \boldsymbol{x}_d = (I - A/\xi_d)^{-1} A\boldsymbol{v}_1 \tag{6.10}$$

in parallel, one on each processor, using the first basis vector $\boldsymbol{v}_1$ as continuation vector. This immediately gives d new basis vectors, which can then be orthogonalized by a master processor. Unfortunately, these vectors tend to become linearly dependent and this approach may easily become unstable as d gets larger. Before demonstrating this, let us recall the role of the upper triangular matrix $U_m = [u_{i,j}]$ introduced in Section 5.1, which collects the coordinates of the continuation vectors $\boldsymbol{y}_j$ (cf. (5.1))

$$\boldsymbol{x}_j = (I - A/\xi_j)^{-1} A\boldsymbol{y}_j, \quad \boldsymbol{y}_j = \sum_{i=1}^{j} \boldsymbol{v}_i u_{i,j}.$$

Obviously, the matrix U_m holds the information as to which of the already computed Krylov basis vectors $\boldsymbol{v}_i$ are used for computing the vector $\boldsymbol{x}_j$. Different matrices U_m yield different sorts of parallelism through partial fractions. For example, with $U_m = I_m$ we obtain a sequential rational Arnoldi algorithm because the system for $\boldsymbol{x}_j$ can only be solved if $\boldsymbol{v}_j$ is available. In the left column of Figure 6.2, labeled as variant (a), we illustrate this situation. In the top picture we show the nonzero pattern of U_m and below we illustrate the entries of the Hessenberg matrix $\underline{H_m}$ computed by the rational Arnoldi algorithm using A, $\boldsymbol{b}$ and ξ_j given in (6.9).

The variant (b) in Figure 6.2 corresponds to the above mentioned extreme case (6.10): here we compute $d = 40$ basis vectors of a rational Krylov space in parallel by solving the shifted linear systems $(I - A/\xi_j)\boldsymbol{x}_j = A\boldsymbol{v}_1$ simultaneously (for $j = 1, \ldots, 40$). Note that this is only possible because all shifts ξ_j are pairwise distinct. We observe that the subdiagonal entries of $\underline{H_m}$ decay rapidly. In fact, it is remarkable that all vectors $\boldsymbol{x}_j = (I - A/\xi_j)^{-1} A\boldsymbol{v}_1$ have large components in only a few of the first Krylov basis vectors, say, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{16}$. We hence expect that the rational Krylov space $\mathcal{Q}_{16} = \operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{16}\}$ contains a good approximation for $f^{\xi_j}(A)\boldsymbol{v}_1$ with the function $f^{\tau}(z) = (1 - z/\tau)^{-1} z$, even if $j \geq 16$ such that the pole ξ_j is not contained in $\mathcal{Q}_{16}$. This expectation can be justified theoretically,

and we will consider such rational approximation problems in Chapter 7, particularly in Section 7.5.

With variant (c) in Figure 6.2 we first compute sequentially 8 Krylov basis vectors, followed by a parallel computation of $d = 8$ Krylov basis vectors at a time, using $\boldsymbol{y}_j = \boldsymbol{v}_{j-8}$ as the continuation vector. With variant (d) we always compute $d = 8$ Krylov basis vectors at a time, using the same continuation vector $\boldsymbol{y}_j = \boldsymbol{v}_{\lceil j/8 \rceil}$ for all of them.

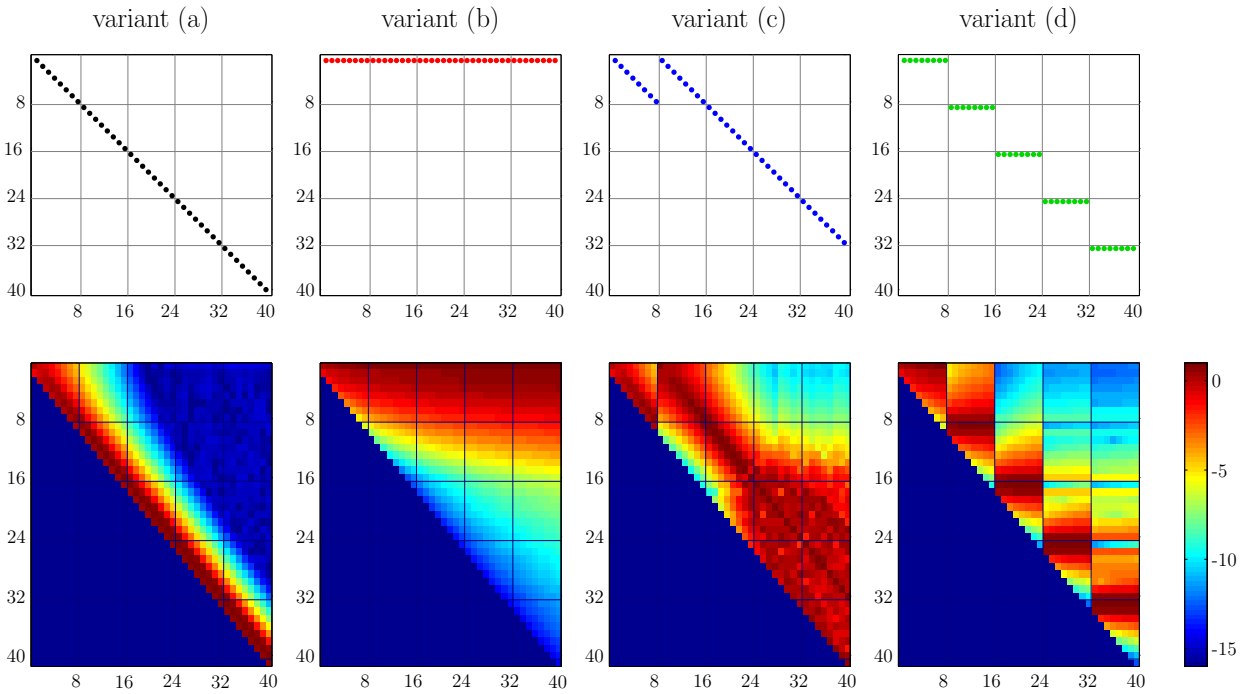


Figure 6.2: In the top row we show the nonzero patterns of four different matrices U_m, $m = 40$. Below are the entries of the Hessenberg matrices $\underline{H_m}$ generated by the rational Arnoldi algorithm with input data (6.9). Dark blue regions correspond to zeros or very small entries of order about 10^{-16}, dark red regions correspond to large entries of order about 10^{1}. Variant (a) is a sequential rational Arnoldi algorithm, and the variants (b)–(c) correspond to parallel rational Arnoldi algorithms.

In Figure 6.3 we show the error curves of Rayleigh approximations $\widehat{\boldsymbol{f}}_m = \widehat{V}_m f(\widehat{A}_m)\widehat{V}_m^* \boldsymbol{b}$, $f(z) = \exp(z)$, computed with the variants (a)–(d) of the rational Arnoldi algorithm (from here on the variables with hats are possibly affected by rounding errors). The error of the projection of $f(A)\boldsymbol{b}$ onto the space $\mathcal{R}(\widehat{V}_m)$ is also shown. The computations were carried out with full reorthogonalization of the basis $\widehat{V}_m$ and the Rayleigh quotient $\widehat{A}_m = \widehat{V}_m^* A \widehat{V}_m$ was computed by explicit projection. All entries of the solution vectors $\boldsymbol{x}_j$ of the linear systems involved are exact to all digits (A is a diagonal matrix and a linear system solve requires exactly one floating point operation per entry of the solution vector). Nevertheless

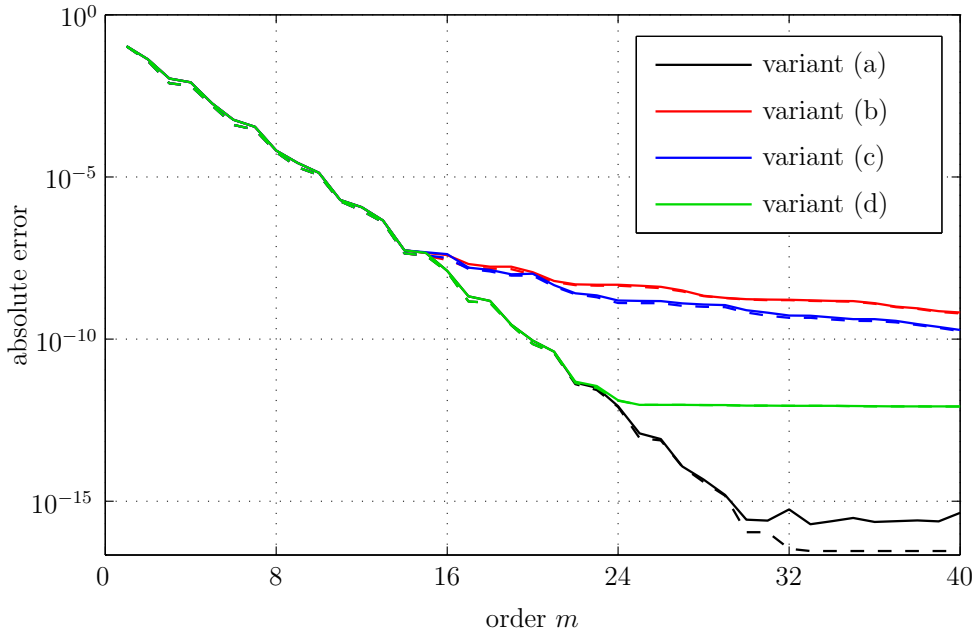


Figure 6.3: Error curves (solid lines) of the Rayleigh approximations $\widehat{\boldsymbol{f}}_m = \widehat{V}_m f(\widehat{A}_m)\widehat{V}_m^* \boldsymbol{b}$ for $f(A)\boldsymbol{b}$, $f(z) = \exp(z)$, computed with the variants (a)–(d) of the rational Arnoldi algorithm, and the error of the projection of $f(A)\boldsymbol{b}$ onto $\mathcal{R}(\widehat{V}_m)$ (dashed lines, partially indistinguishable from the solid lines).


6parallel

we observe that the error curves of variants (b)–(d) stagnate at a higher level than the error curve of the sequential rational Arnoldi algorithm, variant (a). Looking at the errors of the projections we see that this stagnation is obviously caused by the search space $\mathcal{R}(\widehat{V}_m)$, which stops improving at a certain order. This loss of accuracy is reflected neither in the residual norms $\|A\widehat{V}_{m+1}\underline{\widehat{K}_m} - \widehat{V}_{m+1}\underline{\widehat{H}_m}\|$ of the rational Arnoldi decompositions, nor in the distance to orthogonality $\|\widehat{V}_{m+1}^*\widehat{V}_{m+1} - I_{m+1}\|$, see Table 6.1.

	variant (a)	variant (b)	variant (c)	variant (d)
$\|A\widehat{V}_{m+1}\underline{\widehat{K}_m} - \widehat{V}_{m+1}\underline{\widehat{H}_m}\|$	2.79e-14	2.34e-14	2.32e-14	2.56e-14
$\|\widehat{V}_{m+1}^*\widehat{V}_{m+1} - I_{m+1}\|$	5.43e-16	2.19e-15	6.31e-16	5.87e-16

Table 6.1: Residual norms and distance to orthogonality of the decompositions computed by the four variants of the rational Arnoldi algorithm ($m = 40$). These quantities do not indicate any accuracy loss.

The difference between all four variants becomes visible by interpreting the rational Arnoldi algorithm as a modified Gram–Schmidt orthogonalization of

$$X_{m+1} := [\hat{\boldsymbol{v}}_1, \widehat{V}_{m+1}\underline{\widehat{H}_m}] = [\boldsymbol{v}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m].$$

Let $\operatorname{cond}(X_{m+1})$ denote the *2-norm condition number* of X_{m+1}, and let ϵ be the *floating point relative accuracy* of a given finite-precision arithmetic. The numerical behavior of the modified Gram–Schmidt algorithm applied to a well-conditioned matrix is essentially understood, see, e.g., [Bjö67, Ruh83, BP92, GRS97, GLR05] and [Hig02, §19.8]. The case in which this algorithm is used with exactly one reorthogonalization is analyzed in [GLR02]. It is shown that if $X_{m+1} \in \mathbb{R}^{N\times(m+1)}$ is *numerically nonsingular*, that is, $p(N,m)\epsilon \operatorname{cond}(X_{m+1}) < 1$ for a low degree polynomial p in N and m, then

$$X_{m+1} + E_m = \widehat{V}_{m+1}[\boldsymbol{e}_1, \underline{\widehat{H}_m}], \quad \text{where} \quad \|E_m\| \le m^2\epsilon\|X_{m+1}\|.$$

This means that the modified Gram–Schmidt algorithm with reorthogonalization computes a QR decomposition of a nearby matrix. Unfortunately, we can scarcely make use of these bounds because, as shown in Figure 6.4, X_{m+1} is far from being numerically nonsingular even for moderate orders m. Hence, the norm of the perturbation E_m may be large and the space $\mathcal{R}(\widehat{V}_{m+1}) = \mathcal{R}(X_{m+1} + E_m)$ may be "far away" from the rational Krylov space $\mathcal{Q}_{m+1} = \mathcal{R}(X_{m+1})$ of which we aimed to compute an orthonormal basis. We observe in Figure 6.4 that $\epsilon \operatorname{cond}(X_{m+1})$ increases particularly fast for the parallel variants (b)–(d) of the rational Arnoldi algorithm, and that it stagnates at a certain level above one. It seems that, at least in this example, this stagnation happens approximately at the same order m for which the Rayleigh approximations $\widehat{\boldsymbol{f}}_m$ stagnate (cf. Figure 6.3).

In [Sko98] is remarked that possible instabilities in the parallel rational Krylov algorithm for computing eigenvalues of generalized eigenproblems are also reflected in rapidly decaying singular values in the matrices $\underline{\widehat{H}_m}$ and $\underline{\widehat{K}_m}$. Indeed we have

$$\operatorname{cond}(X_{m+1}) = \operatorname{cond}(V_{m+1}[\boldsymbol{e}_1, \underline{H_m}]) = \operatorname{cond}([\boldsymbol{e}_1, \underline{H_m}]),$$

and the last quantity can easily be monitored during the rational Arnoldi algorithm. Note that an almost singular matrix $\widehat{K}_m$ is useless for computing the Rayleigh quotient $\widetilde{A}_m = \widehat{H}_m\widehat{K}_m^{-1}$ by the "last pole at infinity" technique we described in Section 6.1. Unfortunately, even more complications concerning instabilities are expected when using the classical Gram–Schmidt orthogonalization instead of the modified variant, the former being better suited for parallel implementation [GLR05].

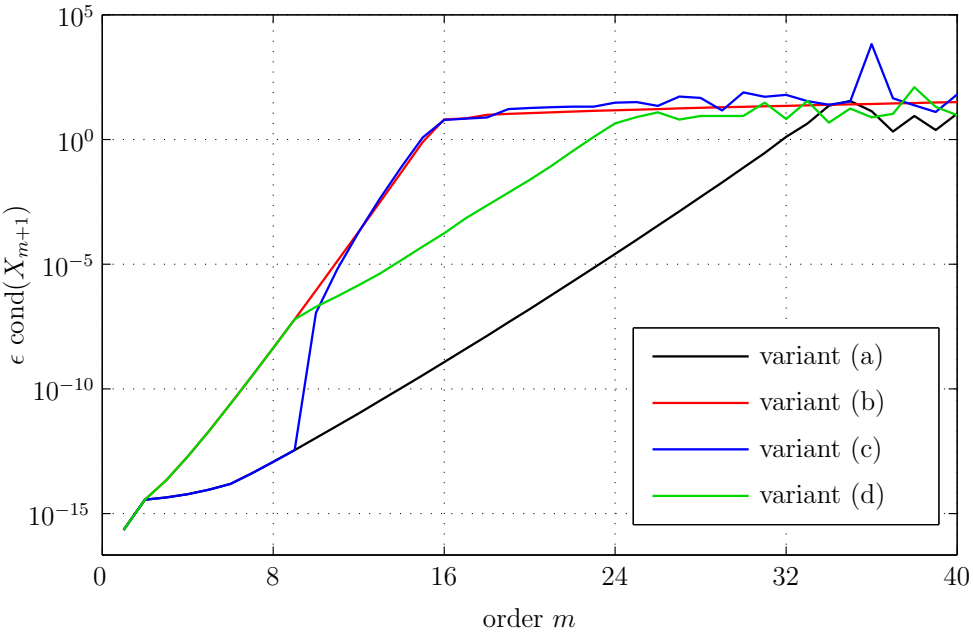


Figure 6.4: The matrix X_{m+1} becomes rapidly ill-conditioned, in particular for the parallel variants (b)–(d) of the rational Arnoldi algorithm.

In summary we have a very interesting situation here: the advantageous approximation properties of rational functions limit the granularity of the parallel rational Arnoldi algorithm because d basis vectors generated in parallel tend to become more and more linearly dependent as d is increased. This problem can be reduced by increasing s instead, i.e., allowing for more parallelism at the linear solver level. The extreme case would be $s = p$, and if all shifts ξ_j were equal we would obtain a parallel variant of the shift-and-invert Arnoldi algorithm, which has been implemented in the "Parallel ARnoldi PACKage" ("shift and invert spectral transformation mode", cf. [MS96, LSY98]). In our numerical experiments we often found $d = 4$ to be a good compromise between slight loss of accuracy and satisfactory parallelization. It may also be advisable to reorder the poles ξ_j to maximize the distance between d consecutive poles.

6.6 A-Posteriori Error Estimation

As a starting point we consider a reduced rational Arnoldi decomposition

$$AV_mK_m = V_mH_m + \boldsymbol{v}_{m+1}\boldsymbol{h}_m^T, \tag{6.11}$$

where $V_{m+1} = [V_m, \boldsymbol{v}_{m+1}]$ has orthonormal columns such that $\mathcal{R}(V_m) = \mathcal{Q}_m$, $\mathcal{R}(V_{m+1}) = \mathcal{Q}_{m+1}$, K_m and H_m are $m \times m$ matrices, H_m is of rank m, and $\boldsymbol{h}_m^T \in \mathbb{C}^{1\times m}$. For simplicity we also assume that $\|\boldsymbol{b}\| = 1$ and $V_m^*\boldsymbol{b} = \boldsymbol{e}_1$. Our aim is to bound or estimate the error $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$ of the Rayleigh–Ritz approximation $\boldsymbol{f}_m = V_m f(A_m)\boldsymbol{e}_1$, $A_m = V_m^*AV_m$. In the literature, a-posteriori error estimates have been derived for polynomial Krylov methods [Saa92a, PS93, ITS10, AEEG08b, DMR08] and for the shift-and-invert method [EH06, Mor07, MN08, Mor09]. We will see that some of these ideas generalize nicely to rational Rayleigh–Ritz approximations.

6.6.1 Norm of Correction

An observation often made in practice is that the error $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$ is of the same order as the norm of the correction $\|\boldsymbol{f}_{m+1} - \boldsymbol{f}_m\|$. By the triangle inequality we have

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \le \|f(A)\boldsymbol{b} - \boldsymbol{f}_{m+1}\| + \|\boldsymbol{f}_{m+1} - \boldsymbol{f}_m\|$$

and if $\|f(A)\boldsymbol{b} - \boldsymbol{f}_{m+1}\|$ is comparably small, the estimate

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \approx \|\boldsymbol{f}_{m+1} - \boldsymbol{f}_m\| \tag{6.12}$$

could be justified. It may fail, e.g., if the approximations $\boldsymbol{f}_m$ stagnate for two or more consecutive iterations.

6.6.2 Cauchy Integral Formula

Using Lemma 5.6 we can transform the decomposition (6.11) into

$$AV_m = V_m A_m + \boldsymbol{v}_{m+1}\boldsymbol{a}_m^T, \tag{6.13}$$

where $A_m = H_m K_m^{-1}$ is the Rayleigh quotient and $\boldsymbol{a}_m^T = \boldsymbol{h}_m^T K_m^{-1} \in \mathbb{C}^{1\times m}$. For this decomposition we can apply a technique used in [ITS10]. Assume that f is analytic in a neighborhood of $\mathbb{W}(A)$ such that Cauchy's integral formula is applicable to define $f(A)$ and $f(A_m)$. With a suitable integration contour Γ we have

$$\begin{aligned} f(A)\boldsymbol{b} - \boldsymbol{f}_m &= f(A)\boldsymbol{b} - V_m f(A_m)\boldsymbol{e}_1 \\ &= \frac{1}{2\pi i}\int_\Gamma f(\zeta)[(\zeta I - A)^{-1}\boldsymbol{b} - V_m(\zeta I_m - A_m)^{-1}\boldsymbol{e}_1]\,\mathrm{d}\zeta. \end{aligned}$$

We write

$$(\zeta I - A)^{-1}\boldsymbol{b} - V_m(\zeta I_m - A_m)^{-1}\boldsymbol{e}_1 = (\zeta I - A)^{-1}\boldsymbol{r}_m(\zeta),$$

where $\boldsymbol{r}_m(\zeta)$ can be interpreted as the residual vector for the approximate solution $\boldsymbol{x}_m = V_m(\zeta I_m - A_m)^{-1}\boldsymbol{e}_1$ of the linear system $(\zeta I - A)\boldsymbol{x} = \boldsymbol{b}$. By (6.13) we have $(\zeta I - A)V_m = V_m(\zeta I_m - A_m) - \boldsymbol{v}_{m+1}\boldsymbol{a}_m^T$, so that

$$\begin{aligned} \boldsymbol{r}_m(\zeta) &= \boldsymbol{b} - [V_m(\zeta I_m - A_m) - \boldsymbol{v}_{m+1}\boldsymbol{a}_m^T](\zeta I_m - A_m)^{-1}\boldsymbol{e}_1 \\ &= \boldsymbol{b} - V_m\boldsymbol{e}_1 + \boldsymbol{v}_{m+1}[\boldsymbol{a}_m^T(\zeta I_m - A_m)^{-1}\boldsymbol{e}_1] \\ &:= \rho_m(\zeta)\boldsymbol{v}_{m+1}, \end{aligned}$$

where $\rho_m(\zeta) = \boldsymbol{a}_m^T(\zeta I_m - A_m)^{-1}\boldsymbol{e}_1$ is a rational function with poles at the rational Ritz values $\Lambda(A_m) = \{\theta_1, \ldots, \theta_m\}$. Assume now that $A_m = X_m D_m X_m^{-1}$ is diagonalizable, i.e., $X_m \in \mathbb{C}^{m\times m}$ is an invertible matrix and $D_m = \mathrm{diag}(\theta_1, \ldots, \theta_m)$. Defining the vectors $[\alpha_1, \ldots, \alpha_m] := \boldsymbol{a}_m^T X_m$ and $[\beta_1, \ldots, \beta_m]^T := X_m^{-1}\boldsymbol{e}_1$, we have

$$\rho_m(\zeta) = \boldsymbol{a}_m^T X_m(\zeta I_m - D_m)^{-1}X_m^{-1}\boldsymbol{e}_1 = \sum_{j=1}^m \alpha_j\beta_j\frac{1}{\zeta - \theta_j}.$$

Now,

$$\begin{aligned} f(A)\boldsymbol{b} - \boldsymbol{f}_m &= \frac{1}{2\pi i}\int_\Gamma f(\zeta)(\zeta I - A)^{-1}\boldsymbol{r}_m(\zeta)\,\mathrm{d}\zeta \\ &= \frac{1}{2\pi i}\int_\Gamma f(\zeta)(\zeta I - A)^{-1}\rho_m(\zeta)\boldsymbol{v}_{m+1}\,\mathrm{d}\zeta \\ &= \sum_{j=1}^{m}\alpha_j\beta_j\frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta-\theta_j}(\zeta I - A)^{-1}\boldsymbol{v}_{m+1}\,\mathrm{d}\zeta \\ &= \sum_{j=1}^{m}\alpha_j\beta_j(f(A) - f(\theta_j)I)(A-\theta_j I)^{-1}\boldsymbol{v}_{m+1}, \end{aligned}$$

where we have used the residue theorem (cf. [Hen88, Thm. 4.7a]) for the last equality. Note that this is an explicit formula for the approximation error, though it involves the term $f(A)$. Following [ITS10] we define the function

$$g_m(\zeta) := \sum_{j=1}^{m}\alpha_j\beta_j \begin{cases} \frac{f(\zeta)-f(\theta_j)}{\zeta-\theta_j}, & \text{if } \zeta \neq \theta_j; \\ f'(\theta_j), & \text{if } \zeta = \theta_j, \end{cases} \tag{6.14}$$

so that for a self-adjoint operator with $\Lambda(A) \subseteq [a,b]$ one can bound the error as

$$\min_{\zeta\in[a,b]}|g_m(\zeta)| \leq \|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq \max_{\zeta\in[a,b]}|g_m(\zeta)|. \tag{6.15}$$

If A is not self-adjoint we still get an upper bound by Crouzeix's theorem (cf. Theorem 4.9)

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq C\max_{\zeta\in\mathbb{W}(A)}|g_m(\zeta)|, \quad C \leq 11.08. \tag{6.16}$$

6.6.3 Auxiliary Interpolation Nodes

The following error indicator is based on an idea described in [Saa92a, Thm 5.1], and extended in [PS93, Thm. 3.1] and [AEEG08b, Sec. 4]. The approach relies on an expansion of the approximation error obtained by adjoining auxiliary nodes $\vartheta_1, \ldots, \vartheta_\ell$ to the interpolation nodes of the rational function underlying $\boldsymbol{f}_m$. We define the nodal polynomials

$$w_0(z) := 1, \qquad w_j(z) := (z-\vartheta_1)\cdots(z-\vartheta_j), \quad j = 1, \ldots, \ell,$$

and note that we have $\boldsymbol{v}_{m+1} = w_0(A)\boldsymbol{v}_{m+1}$ and $w_j(A)\boldsymbol{b} = (A-\theta_j I)w_{j-1}(A)\boldsymbol{v}_{m+1}$. The decomposition (6.11) can obviously be extended to

$$A[V_m, W_{\ell-1}]\left[\begin{array}{c|ccccc} K_m & & & & & \\ \hline & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 \end{array}\right] = [V_m, W_\ell]\left[\begin{array}{c|ccccc} H_m & & & & & \\ \hline \boldsymbol{h}_m^T & \vartheta_1 & & & & \\ & 1 & \vartheta_2 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & \vartheta_\ell \\ & & & & 1 \end{array}\right], \qquad (6.17)$$

where $W_j := [w_0(A)\boldsymbol{v}_{m+1}, w_1(A)\boldsymbol{v}_{m+1}, \ldots, w_j(A)\boldsymbol{v}_{m+1}]$. This is a reduced rational Krylov decomposition and the matrix

$$B_m^\ell := \left[\begin{array}{c|ccccc} H_m & & & & & \\ \hline \boldsymbol{h}_m^T & \vartheta_1 & & & & \\ & 1 & \vartheta_2 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & \vartheta_\ell \end{array}\right]\left[\begin{array}{c|ccccc} K_m & & & & & \\ \hline & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 \end{array}\right]^{-1} \in \mathbb{C}^{(m+\ell)\times(m+\ell)}$$

has the eigenvalues $\Lambda(B_m^\ell) = \Lambda(H_m K_m^{-1}) \cup \{\vartheta_1, \ldots, \vartheta_\ell\} = \Lambda(A_m) \cup \{\vartheta_1, \ldots, \vartheta_\ell\}$. By Theorem 5.8, the rational Krylov approximation $\boldsymbol{f}_m^\ell$ associated with the decomposition (6.17) satisfies

$$\boldsymbol{f}_m^\ell = [V_m, W_{\ell-1}] f(B_m^\ell)\boldsymbol{e}_1 = r_{m+\ell}(A)\boldsymbol{b},$$

where $r_{m+\ell} \in \mathcal{P}_{m+\ell-1}/q_{m-1}$ interpolates f at the eigenvalues $\Lambda(B_m^\ell)$. We have thus added ℓ interpolation nodes to our rational interpolating function. It is interesting to investigate the matrix $f(B_m^\ell)$ further. In [AEEG08b, Lem. 4.1] is shown that

$$f(B_m^\ell) = \left[\begin{array}{c|cccc} f(A_m) & & & & \\ \hline \boldsymbol{a}_m^T\phi_1(A_m) & f(\vartheta_1) & & & \\ \boldsymbol{a}_m^T\phi_2(A_m) & \Delta_1^1 & f(\vartheta_2) & & \\ \vdots & \vdots & \vdots & \ddots & \\ \boldsymbol{a}_m^T\phi_\ell(A_m) & \Delta_1^{\ell-1} & \Delta_2^{\ell-2} & \ldots & f(\vartheta_\ell) \end{array}\right]$$

with the functions

$$\phi_0(z) := f(z), \qquad \phi_j(z) := \frac{\phi_{j-1}(z) - \phi_{j-1}(\vartheta_j)}{z - \vartheta_j}, \quad j = 1, \ldots, \ell,$$

and the kth order divided differences of f with respect to $\vartheta_j, \ldots, \vartheta_{j+k}$ (cf. [Wal69, §3.2])

$$\mathbb{\Delta}_j^k := \frac{1}{2\pi i} \int_\Gamma \frac{f(\zeta)}{(\zeta - \vartheta_j) \cdots (\zeta - \vartheta_{j+k})} \, \mathrm{d}\zeta.$$

The rational Krylov approximation $\boldsymbol{f}_m^\ell$ can now be written as

$$\begin{aligned} \boldsymbol{f}_m^\ell &= [V_m, W_{\ell-1}] f(B_m^\ell) \boldsymbol{e}_1 \\ &= V_m f(A_m) \boldsymbol{e}_1 + \sum_{j=1}^{\ell} (\boldsymbol{a}_m^T \phi_j(A_m) \boldsymbol{e}_1) w_{j-1}(A) \boldsymbol{v}_{m+1} \\ &= \boldsymbol{f}_m + \sum_{j=1}^{\ell} (\boldsymbol{a}_m^T \phi_j(A_m) \boldsymbol{e}_1) w_{j-1}(A) \boldsymbol{v}_{m+1}. \end{aligned}$$

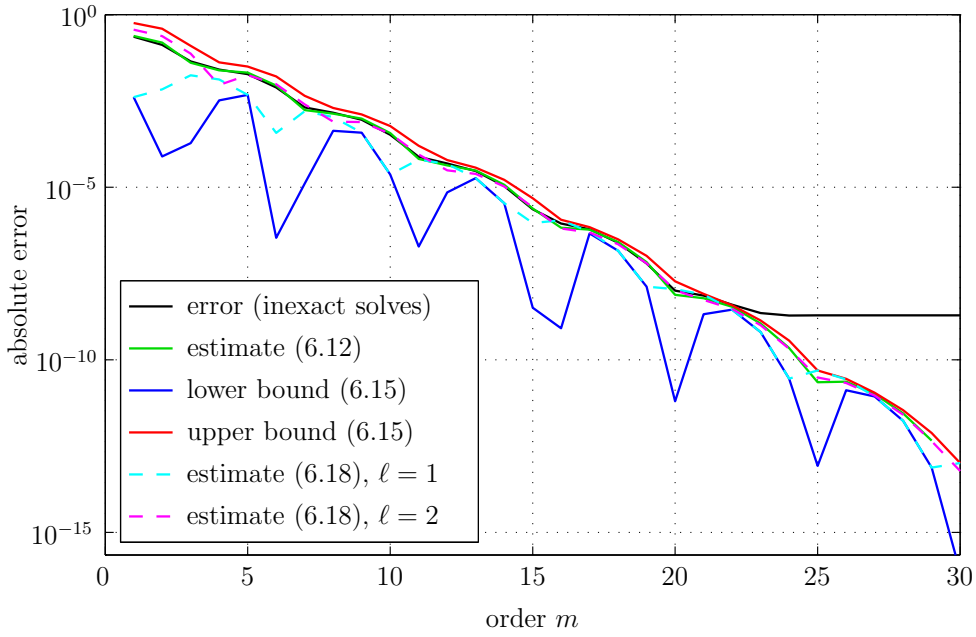
Under the assumption that we can choose a sequence of auxiliary nodes $\{\vartheta_j\}$ such that $\boldsymbol{f}_m^\ell \to f(A)\boldsymbol{b}$ as $\ell \to \infty$, we arrive at the error representation

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| = \Big\| \sum_{j=1}^{\infty} (\boldsymbol{a}_m^T \phi_j(A_m) \boldsymbol{e}_1) w_{j-1}(A) \boldsymbol{v}_{m+1} \Big\|.$$

A practical error estimate is obtained by truncating again the infinite sum to ℓ terms, i.e.,

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \approx \Big\| \sum_{j=1}^{\ell} (\boldsymbol{a}_m^T \phi_j(A_m) \boldsymbol{e}_1) w_{j-1}(A) \boldsymbol{v}_{m+1} \Big\|. \tag{6.18}$$

Example 6.4. For illustration we use the data (6.9) to approximate $f(A)\boldsymbol{b}$, $f(z) = \exp(z)$. In Figure (6.5) we show the error curve of Rayleigh approximations $\boldsymbol{f}_m$ computed with the rational Arnoldi algorithm (black curve, all linear systems were solved to a residual norm of 10^{-8}), and the presented error estimates and bounds. The estimate (6.18) was used with $\ell = 1$ and $\ell = 2$, where $\vartheta_1 = \lambda_{\min} = -99$ and $\vartheta_2 = \lambda_{\max} = 0$ (this choice yields lower and upper bounds in the polynomial Krylov case, see [AEEG08b]). Except for the lower bound (6.15), which shows a quite erratic behavior, all estimates predict well the actual error $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$ until its stagnation due to the inexact linear system solves. In conjunction with the estimate (6.8) for the sensitivity error we thus obtain practical stopping criteria for the rational Arnoldi algorithm.



6error

Figure 6.5: Actual error curve (black) and the error estimates/bounds for approximating $f(A)\boldsymbol{b}$, $f(z) = \exp(z)$, by the Rayleigh–Ritz method. The linear systems involved are solved inexactly, which causes the early stagnation of the error curve.

Remark 6.5. It is interesting to note that for $\ell = 1$ we have

$$\|(\boldsymbol{a}_m^T \phi_1(A_m)\boldsymbol{e}_1) w_0(A)\boldsymbol{v}_{m+1}\| = |\boldsymbol{a}_m^T (f(A_m) - f(\vartheta_j) I_m)(A_m - \vartheta_j I_m)^{-1} \boldsymbol{e}_1|,$$

and, if $A_m = X_m D_m X_m^{-1}$ is diagonalizable, this equals $|g_m(\vartheta_j)|$ with the function g_m defined in (6.14). This explains why the bounds (6.15) obtained by minimizing or maximizing $|g_m(\zeta)|$ among all $\zeta \in \mathbb{W}(A)$ are always below or above the estimate (6.18), respectively, if the auxiliary node ϑ_1 is contained in $\mathbb{W}(A)$.

7 Selected Approximation Problems

In the study of expansions
of analytic functions it is
often of great convenience to
study first the function $1/(t-z)$.
J. L. Walsh [Wal69, §3.1]

In the previous chapters we have considered various methods to extract an approximation $\boldsymbol{f}_m \approx f(A)\boldsymbol{b}$ from a rational Krylov space $\mathcal{Q}_m(A,\boldsymbol{b}) = q_{m-1}(A)^{-1}\mathcal{K}_m(A,\boldsymbol{b})$. All these methods require a choice of parameters:

— The Rayleigh–Ritz method (cf. Section 4.2) depends on the poles $\xi_1,\dots,\xi_{m-1}$ of the rational Krylov space, i.e., on the zeros of q_{m-1}.

— The shift-and-invert method (cf. Section 5.4.3) requires the choice of the shift ξ.

— The extraction by rational interpolation (PAIN method, cf. Section 5.4.2) requires the choice of poles $\xi_1,\dots,\xi_{m-1}$ and interpolation nodes $\alpha_1,\dots,\alpha_m$.

— The evaluation of a partial fraction expansion (PFE method, cf. Section 5.4.4) requires the poles $\xi_1,\dots,\xi_{m-1}$ and the residues $\gamma_0,\gamma_1,\dots,\gamma_{m-1}$ of a partial fraction.

This chapter is devoted to the choice of these parameters.

The above methods have in common that the resulting approximations are of the form $\boldsymbol{f}_m = r_m(A)\boldsymbol{b}$ with a rational function $r_m = p_{m-1}/q_{m-1}$ of type $(m-1,m-1)$. By

Crouzeix's theorem (cf. Theorem 4.9) we know that

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq C\|\boldsymbol{b}\| \, \|f - r_m\|_\Sigma,$$

where $C \leq 11.08$ and $\| \cdot \|_\Sigma$ denotes the uniform norm on a set $\Sigma \supseteq \mathbb{W}(A)$. The aim for a smallest possible approximation error $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$ immediately leads us to the problem of rational best uniform approximation of f on Σ. In particular, the Rayleigh–Ritz approximation $\boldsymbol{f}_m$ is near-optimal,

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq 2C\|\boldsymbol{b}\| \inf_{p \in \mathcal{P}_{m-1}} \|f - p/q_{m-1}\|_\Sigma$$

(cf. Theorem 4.10), and hence the minimization of the approximation error reduces to the problem of finding an optimal denominator q_{m-1}, or equivalently, an optimal search space $\mathcal{Q}_m$. Rational Krylov methods with an underlying rational Krylov decomposition are closely related to problems of rational interpolation (cf. Theorem 5.8), a topic we will also include in our considerations.

In many applications, the function $f = f^\tau$ also depends on a parameter τ from a *parameter set* T, and consequently the same is true for the approximations $\boldsymbol{f}_m^\tau \approx f^\tau(A)\boldsymbol{b}$. This needs to be taken into account when optimizing the parameters of a rational Krylov method. Moreover, it is often necessary to restrict the poles ξ_j to a *pole set* Ξ. For example, if complex arithmetic is to be avoided, $\Xi = \mathbb{R} \cup \{\infty\}$ is an appropriate restriction.

In the next two sections we collect tools from the theory of rational approximation and interpolation, followed by a brief overview of logarithmic potential theory. The remaining sections are devoted to various approaches for computing optimal (in a sense to be specified) parameters for a rational Krylov method, depending on the function f^τ and the configuration of Σ, T and Ξ.

7.1 Preliminaries from Rational Approximation Theory

We consider a closed set $\Sigma \subset \mathbb{C}$ and a function f whose domain of definition contains Σ. Often we will require that f is *analytic on* Σ, by which we mean *analytic in an open set* $\Omega \supset \Sigma$. By $\mathcal{R}_{m,n}$ we denote the set of rational functions of type (m, n), that is, the set of quotients p/q with $p \in \mathcal{P}_m$ and $q \in \mathcal{P}_n$. By $\mathcal{R}_{m,n}^{\Xi}$ we mean the set of rational functions of type (m, n) with all poles in a nonempty closed set $\Xi \subseteq \overline{\mathbb{C}}$. The following fundamental theorem about the *possibility* of uniform approximation of f is due to Runge [Run84] (here we essentially use a formulation from [Wal69, §1.6, Thm. 8]).

Theorem 7.1 (Runge)**.** *Let f be analytic on a closed set $\Sigma \subset \mathbb{C}$. Let a set $\Xi \subset \overline{\mathbb{C}} \setminus \Sigma$ contain at least one point in each connected component into which Σ separates the plane. Then for every $\varepsilon > 0$ there exists a rational function $r(z)$ with all poles in Ξ such that $\|f - r\|_\Sigma < \varepsilon$.*

In particular, if $\overline{\mathbb{C}} \setminus \Sigma$ is connected, $r(z)$ can be chosen as a polynomial.

There is a more general version of this theorem, called Mergelyan's theorem, which only requires f to be analytic in $\mathrm{int}(\Sigma)$ and continuous on Σ (cf. [Kra99]). However, Runge's theorem is sufficient for our primary goal, which will be the construction of rational functions with poles in Ξ such that the error $\|f - r\|_\Sigma$ is smaller than a given tolerance ε. Another task is to obtain a smallest possible error $\|f - r^*\|_\Sigma$ with a rational function $r^* \in \mathcal{R}_{m,m}^{\Xi}$. Such a function r^* is called *rational best uniform approximation to f on Σ with all poles in Ξ*. The following theorem guarantees the *existence* of such rational functions (cf. [Wal69, §12.3, Cor. 1 & Cor. 2]).

Theorem 7.2. *Let f be continuous on a compact set $\Sigma \subset \mathbb{C}$ and let $\Xi \subset \overline{\mathbb{C}} \setminus \Sigma$ be closed and nonempty. Then there exists a rational function $r^* \in \mathcal{R}_{m,m}^{\Xi}$ with*

$$\|f - r^*\|_\Sigma = \min_{r \in \mathcal{R}_{m,m}^{\Xi}} \|f - r\|_\Sigma.$$

This theorem does not say anything about uniqueness, and indeed, rational best uniform approximations are in general not unique (see, e.g., [Wal69, §12.4] or [GT83]).

Assume now we have a compact parameter set T and a parameterized function $f^\tau(z)$, $\tau \in T$. If there exists a continuous function g on $T\Sigma = \{\tau z : \tau \in T, z \in \Sigma\}$ such that

$f^\tau(z) = g(\tau z)$, then by Theorem 7.2 and the compactness of T we know that there exists a rational function $r^* \in \mathcal{R}^{\Xi}_{m,m}$, $\Xi \subset \overline{\mathbb{C}} \setminus T\Sigma$ satisfying

$$\max_{\tau\in T} \|f^\tau(z) - r^*(\tau z)\|_{z\in\Sigma} = \min_{r\in\mathcal{R}^{\Xi}_{m,m}} \max_{\tau\in T} \|f^\tau(z) - r(\tau z)\|_{z\in\Sigma}.$$

[A similar statement holds if $f^\tau(z) = g(z+\tau)$.] Note that the poles of $r^*(\tau z)$ [and $r^*(z+\tau)$] depend on the parameter τ. In the context of rational Krylov methods, however, one is usually interested in optimal rational approximations with poles *independent* of τ. In other words, we are looking for a fixed rational function $1/q^* \in \mathcal{R}^{\Xi}_{0,m}$ such that

$$\inf_{p\in\mathcal{P}_m} \sup_{\tau\in T} \|f^\tau - p/q^*\|_\Sigma = \inf_{1/q\in\mathcal{R}^{\Xi}_{0,m}} \inf_{p\in\mathcal{P}_m} \sup_{\tau\in T} \|f^\tau - p/q\|_\Sigma. \tag{7.1}$$

The following theorem assures that this problem has a solution.

Theorem 7.3. *Let f^τ be defined on Σ for all $\tau \in T$ and let $\Xi \subset \overline{\mathbb{C}} \setminus \Sigma$ be closed and nonempty. Then there exists a rational function $1/q^* \in \mathcal{R}^{\Xi}_{0,m}$ satisfying* (7.1).

Proof. Define

$$\rho(1/q) = \inf_{p\in\mathcal{P}_m} \sup_{\tau\in T} \|f^\tau - p/q\|_\Sigma.$$

We can assume that there exists $1/q \in \mathcal{R}^{\Xi}_{0,m}$ such that $\rho(1/q) < \infty$, otherwise both sides of (7.1) attain the value ∞. Let

$$\alpha = \inf_{1/q\in\mathcal{R}^{\Xi}_{0,m}} \rho(1/q) \geq 0$$

and let $\{1/q_k\}$ be a sequence in $\mathcal{R}^{\Xi}_{0,m}$ such that $\lim_{k\to\infty} \rho(1/q_k) = \alpha$. Without loss of generality we may assume that all polynomials q_k are monic. By the assumption that Ξ be closed, we can extract a subsequence from $\{1/q_k\}$ converging uniformly on Σ to a rational function $1/q_\infty \in \mathcal{R}^{\Xi}_{0,m}$. To see that $\rho(1/q_\infty) = \alpha$ it only remains to show that $\rho(1/q)$ is continuous. Assume without loss of generality that $\rho(1/q_1) < \rho(1/q_2)$. For any $\varepsilon > 0$ there exist $\tau_1 \in T$ and $p_1 \in \mathcal{P}_m$ such that $\|f^{\tau_1} - p_1/q_1\|_\Sigma \leq \rho(1/q_1) + \varepsilon$. Then

$$\begin{aligned}
\rho(1/q_2) &\leq \|f^{\tau_1} - p_1/q_2\|_\Sigma \\
&\leq \|f^{\tau_1} - p_1/q_1\|_\Sigma + \|p_1/q_1 - p_1/q_2\|_\Sigma \\
&\leq \rho(1/q_1) + \varepsilon + \|p_1/q_1 - p_1/q_2\|_\Sigma,
\end{aligned}$$

and hence

$$|\rho(1/q_1) - \rho(1/q_2)| \leq \varepsilon + \|p_1\|_\Sigma \, \|1/q_1 - 1/q_2\|_\Sigma,$$

where $\|p_1\|_\Sigma < \infty$. This proves that ρ is continuous. □

We now turn to the question of the *rate* of approximation, i.e., among all $r_m \in \mathcal{R}^\Xi_{m-1,m-1}$ how fast can the error $\|f - r_m\|_\Sigma$ decay as $m \to \infty$? We will see that this question is closely related to the construction of rational functions which are uniformly as small as possible on Σ and as large as possible on Ξ, and such rational functions can be studied with tools from logarithmic potential theory.

7.2 Preliminaries from Logarithmic Potential Theory

For a detailed treatment of the concepts reviewed here we refer to the monographs [Ran95, ST97]. The connections between logarithmic potential theory and rational interpolation and approximation are nicely outlined in [LS06].

The Classical Case. For a compact set $\Sigma \subset \mathbb{C}$, let $\mathcal{M}(\Sigma)$ denote the set of positive Borel measures μ with support $\operatorname{supp}(\mu) \subseteq \Sigma$ and total mass $\|\mu\| = 1$ (also referred to as *probability measures on* Σ). The *logarithmic potential* of a measure μ is defined as

$$U^\mu(z) = \int \log \frac{1}{|z - \zeta|} \, \mathrm{d}\mu(\zeta).$$

This function is superharmonic in $\mathbb{C}$ and harmonic outside $\operatorname{supp}(\mu)$. The *logarithmic energy of* μ is given by

$$I(\mu) = \int U^\mu \, \mathrm{d}\mu = \iint \log \frac{1}{|z - \zeta|} \, \mathrm{d}\mu(\zeta) \, \mathrm{d}\mu(z).$$

The quantity

$$V_\Sigma = \inf_{\mu \in \mathcal{M}(\Sigma)} I(\mu)$$

is called the *logarithmic energy of* Σ. The *logarithmic capacity of* Σ is defined as

$$\operatorname{cap}(\Sigma) = \exp(-V_\Sigma).$$

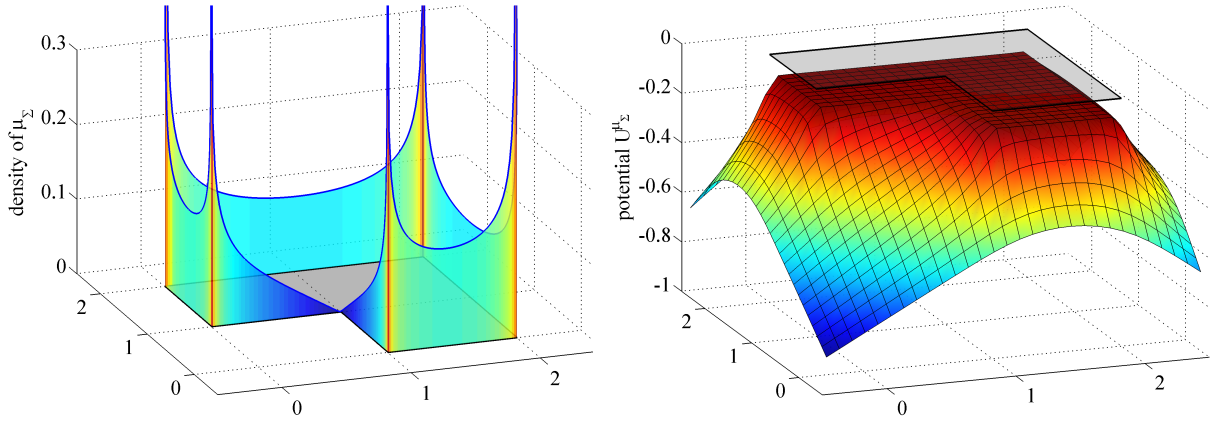


Figure 7.1: The density of the equilibrium measure μ_Σ of an L-shaped domain Σ and the associated potential U^{μ_Σ}. The level of the "plateau" of the potential is $V_\Sigma = -0.082$, so that $\mathrm{cap}(\Sigma) = \exp(-V_\Sigma) = 1.085$. The level lines of the potential are also the level lines of the Green's function g_Ω of the outer domain $\Omega = \mathbb{C} \setminus \Sigma$.

If $V_\Sigma = +\infty$ we set $\mathrm{cap}(\Sigma) = 0$. A property is said to hold *quasi-everywhere (q.e.) on* Σ if it holds everywhere on Σ except for a subset of logarithmic capacity zero. Let us assume that $\mathrm{cap}(\Sigma) > 0$. In this case the fundamental theorem of Frostman asserts that there exists a unique *equilibrium measure* $\mu_\Sigma \in \mathcal{M}(\Sigma)$ such that $I(\mu_\Sigma) = V_\Sigma$ (cf. [Ran95, Thm. 3.3.4]). Let Ω be the *outer domain relative to* Σ, by which is meant the unbounded component of $\mathbb{C} \setminus \Sigma$. Here is a list of remarkable properties of the equilibrium measure (cf. Figure 7.1):

— there holds $\mathrm{supp}(\mu_\Sigma) \subseteq \partial\Omega$, and

— since $\mathcal{M}(\partial\Omega) \subseteq \mathcal{M}(\Sigma)$ and μ_Σ is unique, this inclusion implies $\mathrm{cap}(\Sigma) = \mathrm{cap}(\partial\Omega)$,

— there holds

$$\begin{aligned} U^{\mu_\Sigma}(z) &\leq V_\Sigma \qquad \text{for all } z \in \mathbb{C}, \\ U^{\mu_\Sigma}(z) &= V_\Sigma \qquad \text{for q.e. } z \in \Sigma, \end{aligned}$$

— conversely, if U^μ is constant q.e. on Σ and $I(\mu) < \infty$ for some $\mu \in \mathcal{M}(\Sigma)$, then $\mu = \mu_\Sigma$.

The *Green's function of* Ω *(with pole at* ∞*)* is

$$g_\Omega(z) := -U^{\mu_\Sigma}(z) + V_\Sigma \,.$$

This function is closely related to the growth of polynomials that are uniformly small on Σ, and thereby also related to polynomial interpolation. With the level curves

$$\Gamma_R = \{z \in \Omega \,:\, g_\Omega(z) = \log R\}, \quad R > 1,$$

the following theorem holds (cf. [Wal69, §7.9]).

Theorem 7.4 (Bernstein, Walsh)**.** *Let Σ be a compact set with $\operatorname{cap}(\Sigma) > 0$ and let R be the largest number such that f admits an analytic continuation to $\operatorname{int}(\Gamma_R)$. Then there exists a sequence of* maximally converging *polynomials $\{p_{m-1} \in \mathcal{P}_{m-1}\}_{m\geq 1}$ that interpolate f in a sequence of nodes $\{\alpha_{m,1}, \ldots, \alpha_{m,m}\}_{m\geq 1} \subset \Sigma$ and satisfy*

$$\limsup_{m\to\infty} \|f - p_{m-1}\|_\Sigma^{1/m} = R^{-1}.$$

Let $\{p^*_{m-1} \in \mathcal{P}_{m-1}\}_{m\geq 1}$ be a sequence of polynomials of best uniform approximation to f on Σ. Then

$$\limsup_{m\to\infty} \|f - p^*_{m-1}\|_\Sigma^{1/m} = R^{-1}.$$

Note that for entire functions f, the level R can be chosen arbitrarily large and hence $\|f - p_{m-1}\|_\Sigma$ and $\|f - p^*_{m-1}\|_\Sigma$ decay superlinearly as $m \to \infty$.

Signed Measures. Let $\Sigma \subset \mathbb{C}$ be compact and $\Xi \subset \mathbb{C}$ be closed, both sets of positive capacity and *distance* $\operatorname{dist}(\Sigma, \Xi) := \inf\{|\sigma - \xi| \,:\, \sigma \in \Sigma, \xi \in \Xi\} > 0$. The pair (Σ, Ξ) is called a *condenser* [Bag67, Gon69]. We define the set of signed measures

$$\mathcal{M}(\Sigma, \Xi) = \{\mu = \mu_\Sigma - \mu_\Xi \,:\, \mu_\Sigma \in \mathcal{M}(\Sigma),\ \mu_\Xi \in \mathcal{M}(\Xi)\}$$

and consider the energy problem

$$V = \inf_{\mu \in \mathcal{M}(\Sigma,\Xi)} I(\mu).$$

One can show that V is a positive number and there exists a unique equilibrium measure μ^* such that $I(\mu^*) = V$ (cf. [ST97, Thm. VIII.1.4]). The quantity

$$\operatorname{cap}(\Sigma, \Xi) = 1/V$$

is called the *condenser capacity of* (Σ, Ξ).

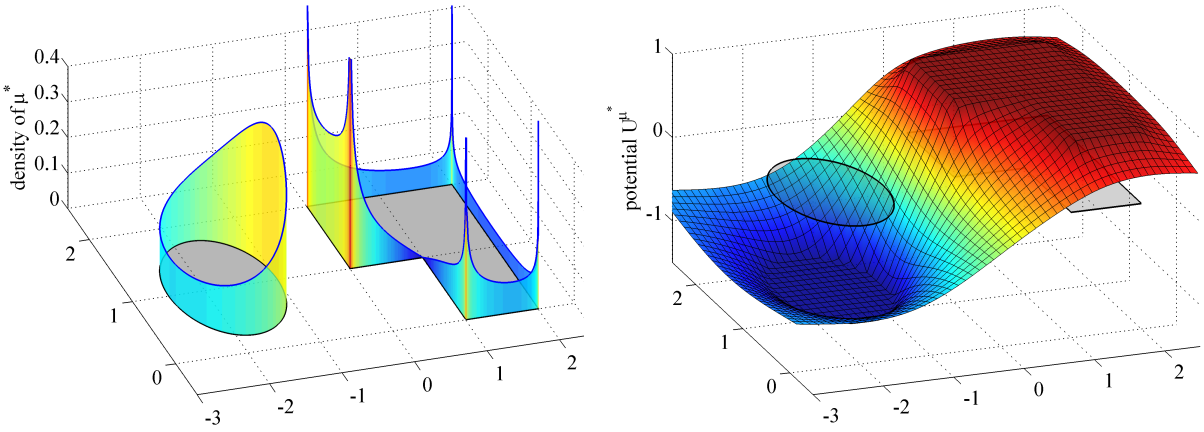


Figure 7.2: The (absolute value of the) density of the signed equilibrium measure μ^* of an L-shaped domain Σ and a circular set Ξ. The associated potential U^{μ^*} is shown on the right. The levels of the "plateaus" of the potential are $F_\Sigma = 0.892$ and $-F_\Xi = -1.174$, so that $\mathrm{cap}(\Sigma, \Xi) = 1/(F_\Sigma + F_\Xi) = 0.484$.

The following assertions hold (cf. Figure 7.2):

— $\mathrm{supp}(\mu^*) \subseteq \partial(\Sigma \cup \Xi)$ (not necessarily the outer boundary),

— $\mathrm{cap}(\partial\Sigma, \partial\Xi) = \mathrm{cap}(\Sigma, \Xi)$,

— there exist real constants F_Σ and F_Ξ such that

$$\begin{aligned} U^{\mu^*}(z) &= F_\Sigma \qquad \text{for q.e. } z \in \Sigma, \\ U^{\mu^*}(z) &= -F_\Xi \qquad \text{for q.e. } z \in \Xi, \end{aligned}$$

— $V = F_\Sigma + F_\Xi$.

To return to problems of rational approximation, let us consider a sequence of rational functions

$$s_m(z) = \frac{(z - \sigma_{m,1}) \cdots (z - \sigma_{m,m})}{(z - \xi_{m,1}) \cdots (z - \xi_{m,m-1})}, \quad m = 1, 2, \ldots \tag{7.2}$$

of type $(m, m-1)$ with zeros $\sigma_{m,j} \in \Sigma$ and poles $\xi_{m,j} \in \Xi$, and the associated *counting measures*

$$\mu_m = \frac{1}{m} \sum_{j=1}^{m} \delta_{\sigma_{m,j}} - \frac{1}{m} \sum_{j=1}^{m-1} \delta_{\xi_{m,j}},$$

where δ_z denotes the Dirac unit measure in the point z. Note that $\mu_m|_\Xi$ is not normalized, but $\|\mu_m|_\Xi\| \to 1$ as $m \to \infty$ (we think of that as if there were an artificial pole $\xi_{m,m} = \infty$).

Now the absolute value of s_m is related to the potential of μ_m by

$$\begin{aligned} U^{\mu_m}(z) &= \frac{1}{m}\sum_{j=1}^{m}\log\frac{1}{|z-\sigma_{m,j}|} - \frac{1}{m}\sum_{j=1}^{m-1}\log\frac{1}{|z-\xi_{m,j}|} \\ &= -\frac{1}{m}\log\Big(\prod_{j=1}^{m}|z-\sigma_{m,j}| \Big/ \prod_{j=1}^{m-1}|z-\xi_{m,j}|\Big) \\ &= -\log|s_m(z)|^{1/m}. \end{aligned}$$

Together with the properties of the equilibrium measure μ^* listed above, the following theorem can be proved [Gon69, LS94].

Theorem 7.5. *For a sequence of rational functions s_m of the form* (7.2) *there holds*

$$\limsup_{m\to\infty}\left(\frac{\sup_{z\in\Sigma}|s_m(z)|}{\inf_{z\in\Xi}|s_m(z)|}\right)^{1/m} \geq e^{-1/\operatorname{cap}(\Sigma,\Xi)}, \tag{7.3}$$

with equality if $\mu_m \xrightarrow{*} \mu^*$.

By $\mu_m \xrightarrow{*} \mu^*$ we mean *weak-star convergence*, i.e., for every continuous function g defined on the supports of all measures μ_m there holds $\int g\,\mathrm{d}\mu_m \to \int g\,\mathrm{d}\mu^*$.

The problem of finding a sequence $\{s_m\}$ such that equality holds in (7.3) is often referred to as *(generalized) Zolotarev problem for the condenser* (Σ,Ξ) because it reduces to the third of Zolotarev's classical problems if Σ and Ξ are real intervals [Zol77, Gon69, Tod84]. Zolotarev problems have been studied extensively in the literature, e.g., in connection with the ADI method [Leb77, EW91, Sta91]. Here are two practical ways (there exist more) to compute zeros and poles of asymptotically optimal rational functions.

— Generalized Fejér points [Wal65]: If Σ and Ξ are closed connected sets (not single points) that do not separate the plane, then by the *Riemann mapping theorem* (cf. [Hen88, Thm. 5.10h]) there exists a function Φ that conformally maps $\Omega = \overline{\mathbb{C}}\setminus(\Sigma\cup\Xi)$ onto a circular annulus $\mathbb{A}_R := \{w : 1 < |w| < R\}$. The number R is called the *Riemann modulus of* $\mathbb{A}_R$ and there holds

$$R = e^{1/\operatorname{cap}(\Sigma,\Xi)}.$$

Denote by $\Psi = \Phi^{-1}$ the inverse map and assume that Ψ can be extended continuously

to $\partial\Sigma$ and $\partial\Xi$ (which is true, e.g., if these boundaries are Jordan curves). The *generalized Fejér points of order m* are then given as

$$\left\{\sigma_{m,j} = \Psi(e^{2\pi ij/m})\right\}_{j=1,\dots,m} \quad \text{and} \quad \left\{\xi_{m,j} = \Psi(Re^{2\pi ij/(m-1)})\right\}_{j=1,\dots,m-1}.$$

— Generalized Leja points [Bag69]: These are nested point sequences, i.e., $\sigma_{m,j} = \sigma_j$ and $\xi_{m,j} = \xi_j$ for $m = 1, 2, \dots$, which is convenient and often crucial for computations. Starting with points $\sigma_1 \in \Sigma$ and $\xi_1 \in \Xi$ of minimal distance, the points $\sigma_{j+1} \in \Sigma$ and $\xi_{j+1} \in \Xi$ are determined recursively such that with

$$s_j(z) = \prod_{i=1}^{j} \frac{z - \sigma_i}{z - \xi_i}$$

the conditions

$$\begin{aligned} \max_{z\in\Sigma} |s_j(z)| &= |s_j(\sigma_{j+1})| \\ \min_{z\in\Xi} |s_j(z)| &= |s_j(\xi_{j+1})| \end{aligned}$$

are satisfied.

By the Walsh–Hermite formula, the error of a rational function r_m with poles $\xi_{m,1}, \dots, \xi_{m,m-1}$ that interpolates f at the nodes $\sigma_{m,1}, \dots, \sigma_{m,m}$ is given by

$$f(z) - r_m(z) = \frac{1}{2\pi i} \int_\Gamma \frac{s_m(z)}{s_m(\zeta)} \frac{f(\zeta)}{\zeta - z} \, \mathrm{d}\zeta, \quad z \in \Sigma,$$

where Γ is a suitable integration contour winding around Σ such that f is analytic in $\mathrm{int}(\Gamma)$ and extends continuously to Γ. The uniform error can be estimated as

$$\|f - r_m\|_\Sigma \le D \frac{\max_{z\in\Sigma} |s_m(z)|}{\min_{\zeta\in\Gamma} |s_m(\zeta)|},$$

where $D = D(f, \Gamma)$ is a constant. In conjunction with Theorem 7.5 we obtain

$$\limsup_{m\to\infty} \|f - r_m\|_\Sigma^{1/m} \le e^{-1/\operatorname{cap}(\Sigma,\Gamma)}, \tag{7.4}$$

provided the nodes $\sigma_{j,m}$ and poles $\xi_{j,m}$ are asympotically distributed according to the equilibrium measure μ^* of the condenser (Σ, Γ). If Σ does not separate the plane (and is

not a single point) then $\mathrm{int}(\Gamma) \setminus \Sigma$ is conformally equivalent to an annulus $\mathbb{A}_R$ and (7.4) can be also written as

$$\limsup_{m\to\infty} \|f - r_m\|_\Sigma^{1/m} \leq R^{-1}. \tag{7.5}$$

By bending the integration contour Γ to enclose a largest possible area $\mathrm{int}(\Gamma)$ (in which the function f is still analytic) we can make $\mathrm{cap}(\Sigma, \Gamma)$ as large as possible (cf. Figure 7.3). Obviously, equilibrium-distributed points on the condenser (Σ, Γ) are reasonable nodes and poles for rational interpolating functions, and we will use these points, e.g., as parameters for the PAIN method (cf. Section 5.4.2).

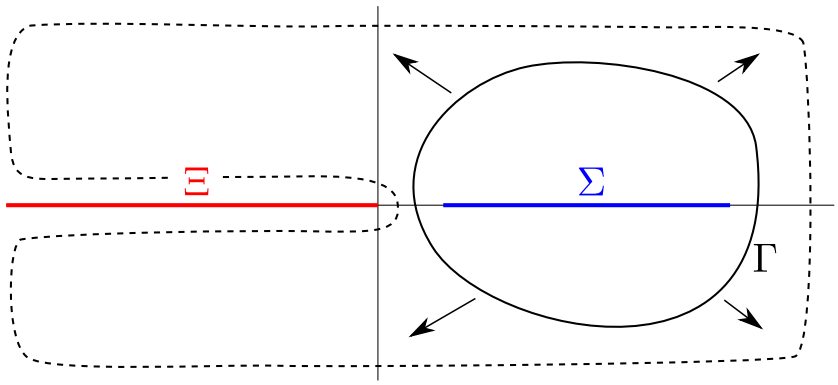


Figure 7.3: Suppose a function f is analytic in the slit plane $\mathbb{C} \setminus \Xi$, $\Xi = (-\infty, 0]$. We can bend the integration contour Γ arbitrarily wide as long as we avoid Ξ, and in the limit the condenser (Σ, Γ) has the same capacity as the condenser (Σ, Ξ).

Remark 7.6. If f is entire we can choose a circle $\Gamma = r\partial\mathbb{D}$ ($\partial\mathbb{D}$ denoting the boundary of the open unit disk $\mathbb{D}$) and let the radius r approach infinity, thus making $\mathrm{cap}(\Sigma, \Gamma)$ arbitrarily large. This means that the poles of the rational interpolating functions move closer and closer to infinity and in the limit we obtain superlinearly converging interpolation polynomials.

Remark 7.7. It is interesting to note that the rational best uniform approximation r_m^* of type $(m-1, m-1)$ to f on Σ is (not better than) twice as good as what we achieved in (7.4). More precisely,

$$\liminf_{m\to\infty} \|f - r_m^*\|_\Sigma^{1/m} \geq e^{-2/\,\mathrm{cap}(\Sigma,\Gamma)} \tag{7.6}$$

(cf. [Par88, Pro93, Pro05]). This result is sharp in the sense that equality holds in (7.6) for certain classes of analytic functions, for example Markov functions [Gon78] or functions with a finite number of algebraic branch points [Sta89].

Remark 7.8. Zolotarev problems can be studied more generally for *ray sequences* $\{r_{m,n} \in \mathcal{R}_{m,n}\}$, that is, $m/n \to \lambda \geq 1$ as $m+n \to \infty$ [LS94]. In a practical rational Krylov algorithm it is usually more time-consuming to expand the Krylov space with a finite pole than with a pole at infinity (which corresponds to a polynomial Krylov step). On the other hand, finite poles usually yield better search spaces $\mathcal{Q}_m$. With the theory of ray sequences it should be possible to study the convergence of a rational Krylov method where a $1/\lambda$-fraction of all poles ξ_j is required to be at infinity. This could then be used to optimize λ for a particular computer if the respective execution times of a rational and a polynomial Krylov step are known.

7.3 The Quadrature Approach

Given a Cauchy integral

$$f(z) = \frac{1}{2\pi i} \int_\Gamma \frac{f(\zeta)}{\zeta - z} \, \mathrm{d}\zeta, \quad z \in \Sigma,$$

the application of a quadrature rule with nodes $\{\xi_j\}_{j=1}^{m-1} \subset \Gamma$ immediately yields a rational function of the form

$$r_m(z) = \sum_{j=1}^{m-1} \frac{w_j}{z - \xi_j} \tag{7.7}$$

with residues w_j. Conversely, rational approximations can usually be interpreted as quadrature rules, the poles being connected by a contour Γ. The use of the trapezoid rule for the approximation of matrix functions is advocated in [ST07b, Sch07, HHT08]. It is known that the convergence of the trapezoidal rule for periodic analytic functions can be very rapid and the rapidity depends on the "thickness" of the region to which the function f can be continued analytically (see [Dav59], [DR84, §4.6.5]). The quadrature approach is even successful for unbounded sets Σ if the function f decays sufficiently fast along the contour Γ, which then passes through infinity (cf. [Tal79, TWS06, ST07a]). Other quadrature concepts, such as Sinc quadrature [Ste93, Ste94, Ste00], have also been applied successfully for operator functions [GHK04].

In [HHT08] the authors consider functions f analytic in the slit plane $\mathbb{C}\backslash(-\infty, 0]$ and a real interval of approximation $\Sigma = [a, b]$, $0 < a < b$. By the Riemann mapping theorem there exists a conformal map Φ that carries the doubly connected region $\Omega = \mathbb{C} \setminus ((-\infty, 0] \cup \Sigma)$

(also known as a *Teichmüller domain* [Ahl73, §4.11]) onto the annulus $\mathbb{A}_R$ with Riemann modulus R. The map Φ can be given explicitly in terms of elliptic functions and

$$R = \exp\left(\frac{\pi}{2}\frac{K(\kappa)}{K(1-\kappa)}\right), \quad \text{where } \kappa = \frac{\sqrt{b/a}-1}{\sqrt{b/a}+1} \tag{7.8}$$

and $K(\kappa) = \int_0^1 [(1-t^2)(1-\kappa t^2)]^{-1/2}\,\mathrm{d}t$ is the complete elliptic integral of the first kind[1].

The application of the trapezoid rule for the transplanted Cauchy integral

$$f(z) = \frac{1}{2\pi i}\int_{\sqrt{R}\partial\mathbb{D}} \frac{f(\Psi(w))}{\Psi(w)-z}\Psi'(w)\,\mathrm{d}w, \quad z \in \Sigma, \tag{7.9}$$

yields a rational function r_m whose poles are the images of $\Psi = \Phi^{-1}$ of $m-1$ equispaced points on the circle $\sqrt{R}\partial\mathbb{D}$ (this corresponds to the "Method 1" in [HHT08]). Indeed, the integrand in (7.9) is a periodic analytic function as w goes along $\sqrt{R}\partial\mathbb{D}$, and results about the asymptotic convergence of the trapezoid rule show that

$$\|f - r_m\| = O(R^{-m/2}). \tag{7.10}$$

This rate can be improved if $(-\infty, 0)$ is just a branch cut of f: applying the transplanted trapezoid rule (7.9) for the function $f(\hat{z})$ in the new variable $\hat{z} = z^{1/2}$, $\hat{z} \in \widehat{\Sigma} = [\sqrt{a}, \sqrt{b}]$, we obtain a convergence of order $O(\widehat{R}^{-m/2})$, where

$$\widehat{R} = \exp\left(\frac{\pi}{2}\frac{K(\hat{\kappa})}{K(1-\hat{\kappa})}\right), \quad \hat{\kappa} = \frac{\sqrt[4]{b/a}-1}{\sqrt[4]{b/a}+1}, \tag{7.11}$$

is the improved Riemann modulus we get when $\mathbb{C} \setminus ((-\infty, 0] \cup \widehat{\Sigma})$ is mapped conformally onto an annulus (this is the "Method 2" in [HHT08]). As the ratio b/a becomes larger, $\widehat{R}$ approaches R^2, and hence the improved quadrature rule converges for large ratios b/a as

$$\|f - r_m\| \approx O((R-\varepsilon)^{-m}),$$

i.e., twice as fast as (7.10). We recall that the same rate $O(R^{-m})$ is achieved by rational interpolation with equilibrium-distributed nodes and poles on the condenser $(\Sigma, (-\infty, 0])$ independently of how large the ratio b/a is, see (7.5). In Figure 7.4 we illustrate the

[1] The definition of $K(\kappa)$ is not consistent in the literature. We stick to the definition in [AS84, §17.3], which is also used by MATLAB when typing `ellipke(kappa)`.

convergence of the Methods 1–3 in [HHT08] and the PAIN method. ("Method 3" is based on the observation that for $f(z) = z^{1/2}$ the function $f(z) = \hat{z}$ has no singularity after the transformation. This method actually reproduces Zolotarev's best relative approximation of the square root.)

The quadrature approach has the feature that it directly yields a rational function in partial fraction form (7.7), which can be evaluated easily in parallel. Moreover, the poles of this rational function do not depend on f, only on the map Φ. If $f = f^\tau$ depends on a parameter $\tau \in T$, one obtains rational functions

$$r_m^\tau(z) = \sum_{j=1}^{m-1} \frac{w_j^\tau}{z - \xi_j} \approx f^\tau(z),$$

where only the parameter-dependent residues w_j^τ need to be recalculated for all $\tau \in T$. At the operator level this means that only $m-1$ linear systems $(A - \xi_j I)^{-1}\boldsymbol{b}$ need to be solved in order to compute $r_m^\tau(A)\boldsymbol{b}$ for various τ.

A drawback of the quadrature approach is that the poles ξ_j are in general complex. This introduces possibly unwanted complex arithmetic in practical algorithms. (The "Method 3" in [HHT08] avoids complex arithmetic, but it is tailored to $f(z) = z^{1/2}$.) If the region Ω is symmetric with respect to the real axis, the poles ξ_j occur in complex conjugate pairs, which can usually be exploited to halve the number of linear system solves.

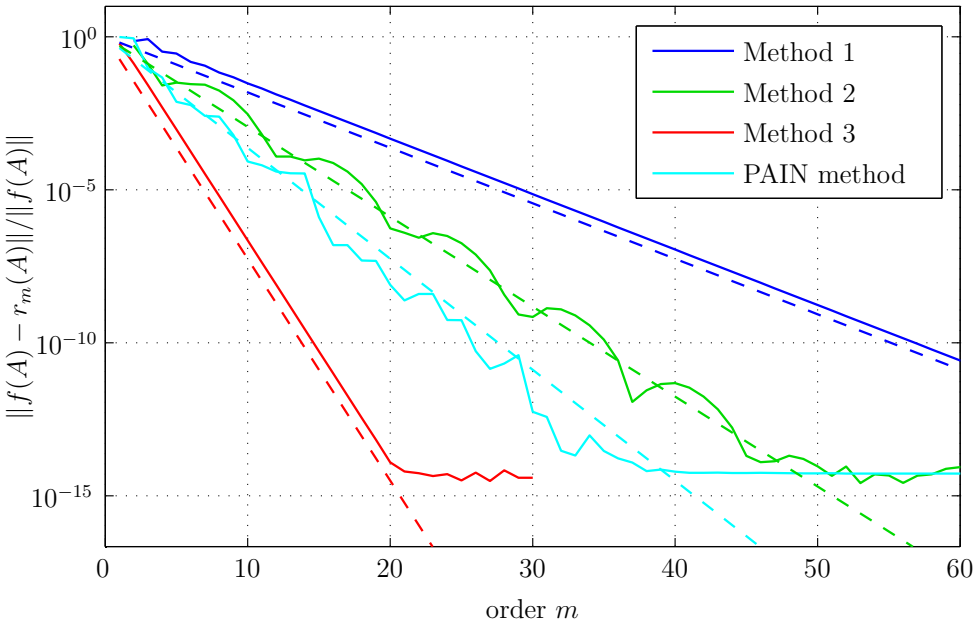


Figure 7.4: Convergence of the Methods 1–3 in [HHT08] for approximating $A^{1/2}$, where the matrix A is generated with the MATLAB command `pascal(5)`. We also show the asymptotic convergence rates $R^{1/2}$, $\widehat{R}$, R^2 (dashed lines), the numbers R and $\widehat{R}$ being given in (7.8) and (7.11), respectively. The "matricized" PAIN method (see Remark 5.10) with generalized Leja points on the sets $\Sigma = [a, b]$ and $\Xi = (-\infty, 0]$ converges at rate R. The plot looks similar for $\log(A)$, except that the red curve needs to be omitted since Method 3 is not applicable in this case.


7quadr

7.4 Single Repeated Poles and Polynomial Approximation

Let us make a general remark about rational Krylov spaces generated with a single repeated pole ξ. Note that every rational function of the form $r_m(z) = p_{m-1}(z)/(z-\xi)^{m-1}$ can be written as a polynomial $\hat{p}_{m-1}(\hat{z})$ in the variable $\hat{z} = 1/(z-\xi)$. Rational approximations of this form are also referred to as *restricted denominator approximations* [Nør78, MN04]. Such approximations are closely related to the shift-and-invert method (cf. Section 5.4.3). We have

$$\|f(z) - r_m(z)\|_\Sigma = \|f(z) - p_{m-1}(z)/(z-\xi)^{m-1}\|_\Sigma = \|\hat{f}(\hat{z}) - \hat{p}_{m-1}(\hat{z})\|_{\widehat{\Sigma}},$$

where $\hat{f}(\hat{z}) := f(\hat{z}^{-1} + \xi)$. Instead considering a rational approximation problem with the function $f(z)$ on the set Σ, we have an equivalent polynomial approximation problem with $\hat{f}(\hat{z})$ on $\widehat{\Sigma} = \{\hat{z} : z \in \Sigma\}$. This simplifies matters in cases where it is possible to solve the polynomial best uniform approximation problem explicitly, or at least approximately (cf. [EH06], where the authors use the Remez algorithm to compute optimal poles for the shift-and-invert method).

Note that even if the closed set Σ is unbounded, the set $\widehat{\Sigma}$ is compact if $\xi \notin \Sigma$. If in addition $\operatorname{cap}(\widehat{\Sigma}) > 0$ and all other requirements of Theorem 7.4 are satisfied, we can characterize the asymptotic convergence of polynomial best uniform approximations utilizing the Green's function $g_{\mathbb{C}\setminus\widehat{\Sigma}}$. In what follows we use this characterization to construct asymptotically best converging rational approximations to the resolvent function.

7.5 The Resolvent Function

In this section we approximate the resolvent function (or transfer function)

$$f^\tau(z) = (z-\tau)^{-1} \tag{7.12}$$

on a closed set Σ by rational functions $r_m^\tau = p_{m-1}^\tau / q_{m-1}$ with poles in a closed set Ξ. The parameters τ are collected in a parameter set T. Since f^τ is a rational function itself, this problem is only interesting if either T has much more than m elements or if $\Xi \neq T$.

As the transfer function is connected with the input-output behavior of linear dynamical systems in the frequency domain, it appears, e.g., in engineering problems, such as filter design or electric circuit simulations, or in geophysical applications. The parameters τ usually correspond to frequencies and are often needed for purely imaginary values only. Polynomial and rational Krylov order reduction techniques for approximating the transfer function over a large frequency interval have proved to be very efficient, especially for large problems [GGV96, RS98, Bai02, Fre03, OR06, BES08, KDZ09].

7.5.1 Single Repeated Poles

Suppose we want to approximate (7.12) on the interval $\Sigma = [a, b]$, $0 \leq a < b$, by a rational function of type $(m-1, m-1)$ with all poles at a point ξ. Using the notation introduced in Section 7.4, the Green's function for $\mathbb{C} \setminus \widehat{\Sigma}$ is

$$g_{\mathbb{C}\setminus\widehat{\Sigma}}(\hat{z}) = \log\left|\zeta + \sqrt{\zeta^2 - 1}\right|, \quad \zeta = \frac{2\hat{z} - (b-\xi)^{-1} - (a-\xi)^{-1}}{(b-\xi)^{-1} - (a-\xi)^{-1}}. \tag{7.13}$$

Note that $\hat{f}^{\tau}(\hat{z})$ has a pole at $\hat{s} = (\tau - \xi)^{-1}$ and therefore $\exp(g_{\mathbb{C}\setminus\widehat{\Sigma}}(\hat{s}))$ is the asymptotic convergence rate we expect from the error of the rational Krylov method with all poles at ξ. It is now easy to consider $R = \exp(g_{\mathbb{C}\setminus\widehat{\Sigma}}(\hat{s}))$ as a function of ξ to find a single repeated asymptotically optimal pole ξ_0.

To obtain transparent formulas, let us assume that the parameters τ are purely imaginary. Then $f^{-\tau}(z) = \overline{f^{\tau}(z)}$ and it suffices to consider a parameter set T on the positive imaginary axis only.

Example 7.9. If $\Sigma = [0, +\infty]$, formula (7.13) evaluated at $\hat{s} = (\tau - \xi)^{-1}$ takes the particularly simple form

$$g_{\mathbb{C}\setminus\widehat{\Sigma}}(\hat{s}) = \log\left|\zeta + \sqrt{\zeta^2 - 1}\right|, \quad \zeta = \frac{\tau + \xi}{\tau - \xi}. \tag{7.14}$$

Let $T = \{\tau\}$ be a singleton parameter set, where $\tau \in i\mathbb{R}_+$ is a positive imaginary number, and let us find an optimal real pole $\xi_0 < 0$. Note that $\zeta = \zeta(\xi)$ as a function of $\xi < 0$ is the Cayley transform and its image U is the upper half of the unit circle as ξ moves along $(-\infty, 0)$. The level lines of $R = \exp(g_{\mathbb{C}\setminus\widehat{\Sigma}}) = \left|\zeta + \sqrt{\zeta^2 - 1}\right|$ as a function of ζ are ellipsoids with foci ± 1. Hence the maximum value of $\exp(g_{\mathbb{C}\setminus\widehat{\Sigma}})$ on U is attained in the point $\zeta_0 = i$,

and this means that $\xi_0 = i\tau$ is the optimal pole. The resulting optimal convergence rate is

$$R = \left|\zeta_0 + \sqrt{\zeta_0^2 - 1}\right| = \left|i + \sqrt{-2}\right| = 1 + \sqrt{2}.$$

More generally, let now $T = [\tau_{\min}, \tau_{\max}]$ be an interval on the positive imaginary axis. Then it is easy to see that $\xi_0 = i\sqrt{\tau_{\min}\tau_{\max}}$ is the optimal pole since

$$\zeta(\tau) = \frac{\tau + \xi_0}{\tau - \xi_0}$$

describes a subarc of U that is symmetric to the imaginary axis as τ moves along T. Note that we get exactly the same subarc of U if we replace T by cT and ξ_0 by $c\xi_0$ for some $c > 0$. Thus U is only dependent on the ratio $c = \tau_{\max}/\tau_{\min}$. In particular,

$$\zeta(\tau_{\min}) = \frac{i - \sqrt{c}}{i + \sqrt{c}}, \quad \zeta(\tau_{\max}) = \frac{\sqrt{c} + i}{\sqrt{c} - i}.$$

Inserting these expressions into the Green's function (7.14), we obtain after some elementary calculations the worst-case convergence rate on T as

$$\begin{aligned} R_1 &= \left|\zeta(\tau_{\min}) + \sqrt{\zeta(\tau_{\min})^2 - 1}\right| = \left|\zeta(\tau_{\max}) + \sqrt{\zeta(\tau_{\max})^2 - 1}\right| \\ &= \left(1 + \frac{\sqrt{8}c^{3/4} + 4c^{1/2} + \sqrt{8}c^{1/4}}{1 + c}\right)^{1/2}. \end{aligned} \tag{7.15}$$

As seen above, the best-case convergence rate $R = 1 + \sqrt{2}$ on T is obtained for the parameter $\tau = \xi_0/i$.

Example 7.10. Let again $\Sigma = [0, +\infty]$ and $T = [\tau_{\min}, \tau_{\max}]$ be an interval on the positive imaginary axis. Let us now choose an *imaginary* optimal pole $\xi_0 \in T$. Note that

$$\zeta(\tau) = \frac{\tau + \xi_0}{\tau - \xi_0}$$

is a real-valued function that should attain values as large as possible for all parameters $\tau \in T$ (i.e., values far away from the interval $[-1, 1]$, of which $g_{\mathbb{C}\setminus\widehat{\Sigma}}$ in (7.14) is the Green's function as a function of ζ). Since $|\zeta(\tau)|$ is monotonically increasing for $\tau \in [\tau_{\min}, \xi_0)$ and monotonically decreasing for $\tau \in (\xi_0, \tau_{\max}]$, we need only consider the endpoints of the

interval T, yielding the equation

$$-\frac{\tau_{\min}+\xi_0}{\tau_{\min}-\xi_0} = \frac{\tau_{\max}+\xi_0}{\tau_{\max}-\xi_0}$$

with solution $\xi_0 = \sqrt{\tau_{\min}\tau_{\max}}$. Note that, as in the previous example, the image of $\zeta(\tau)$ does not change if we replace T by cT and ξ_0 by $c\xi_0$ for some $c > 0$. In particular,

$$\zeta(\tau_{\min}) = \frac{1+\sqrt{c}}{1-\sqrt{c}}, \quad \zeta(\tau_{\max}) = \frac{\sqrt{c}+1}{\sqrt{c}-1}$$

with the ratio $c = \tau_{\max}/\tau_{\min}$.

The best-case convergence rate $R = \infty$ on T is obviously obtained for $\tau = \xi_0$. The worst-case convergence rate on T is

$$\begin{aligned} R_2 &= \left|\zeta(\tau_{\min}) + \sqrt{\zeta(\tau_{\min})^2-1}\right| = \left|\zeta(\tau_{\max}) + \sqrt{\zeta(\tau_{\max})^2-1}\right| \\ &= \frac{c^{1/4}+1}{c^{1/4}-1}\,. \end{aligned} \tag{7.16}$$

Note that the convergence rate (7.16) is always better than what we achieved in (7.15) using a real pole. However, both rates R_1 and R_2 tend to 1 as $c \to \infty$ (cf. Figure 7.5). For a practical implementation of a near-optimal rational Krylov method[2] this implies that it does not pay off to use an imaginary pole (and hence complex arithmetic) if only c is large enough. Assume that one iteration of the rational Krylov method with an imaginary pole takes d times as long as one iteration with a real pole. Then the break-even point of the computation time of both methods satisfies $R_1^{-dm} = R_2^{-m}$, i.e., $d = \log(R_2)/\log(R_1)$. As an example, let $d = 4$. Then Figure 7.5 shows that the method with the real pole will already outperform the method with the imaginary pole if $c \gtrapprox 1.27$.

Remark 7.11. Equation (7.15) can be used to give a formula for the asymptotic convergence rate as a function of $\tau \in i\mathbb{R}_+$ for a fixed pole $\xi_0 < 0$. Note that ξ_0 is the optimal pole for all imaginary parameter intervals $[-\xi_0^2/\tau, \tau]$ with ratio $c = -\tau^2/\xi_0^2$. Since $\tau = \tau_{\max}$ for

[2] Near-optimal means that the approximation $\boldsymbol{f}_m$ extracted from the rational Krylov space $\mathcal{Q}_m$ converges to the exact solution with at least the same asymptotic rate as the best approximation from $\mathcal{Q}_m$ when $m \to \infty$. The Rayleigh–Ritz method is near-optimal.

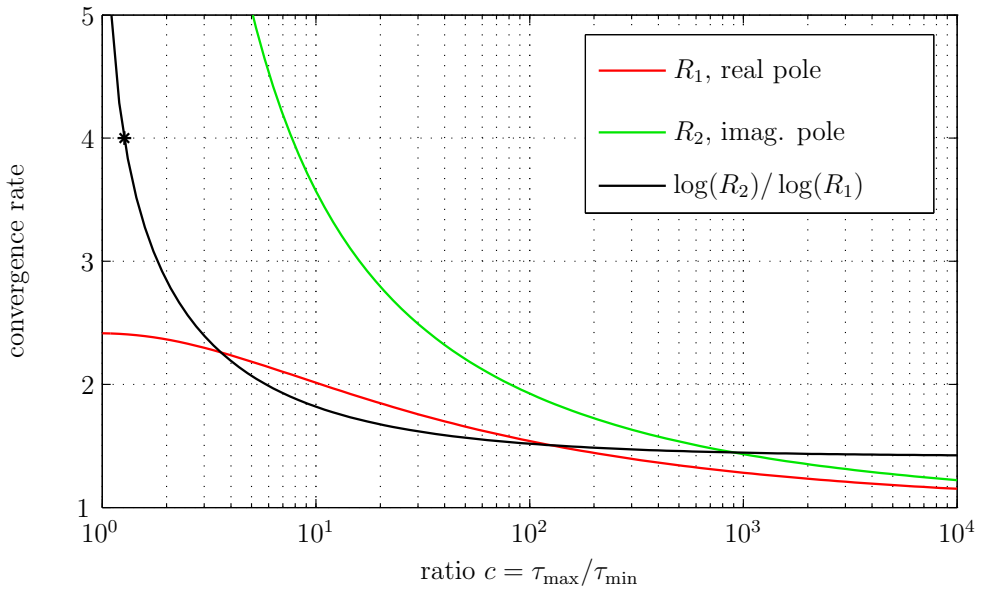


Figure 7.5: Worst-case convergence rates on the parameter set $T = [\tau_{\min}, \tau_{\max}]$ as a function of $c = \tau_{\max}/\tau_{\min}$ using a single repeated optimal real pole (R_1, red) compared to an optimal imaginary pole (R_2, green). The logarithmic quotient (black line) indicates how many iterations are required by the rational Krylov method using the optimal real pole to achieve the same error reduction as one iteration of the method using the optimal imaginary pole.

all these intervals, we can substitute c in (7.15). To summarize,

$$R_1(\tau, \xi_0) = \left(1 + \frac{\sqrt{8}c^{3/4} + 4c^{1/2} + \sqrt{8}c^{1/4}}{1+c}\right)^{1/2}, \quad c = -\tau^2/\xi_0^2.$$

The same reasoning can be applied to (7.16) with a fixed pole $\xi_0 \in i\mathbb{R}_+$, which results in

$$R_2(\tau, \xi_0) = \frac{c^{1/4}+1}{c^{1/4}-1}, \quad c = \tau^2/\xi_0^2.$$

In Figure 7.6 we show a graph of these functions.

7.5.2 Connection to Zolotarev Problems

Given a rational function $s_m \in \mathcal{R}^{\Xi}_{m,m-1}$, it is easily verified that

$$r_m^{\tau}(z) = \frac{1 - \frac{s_m(z)}{s_m(\tau)}}{z - \tau}$$

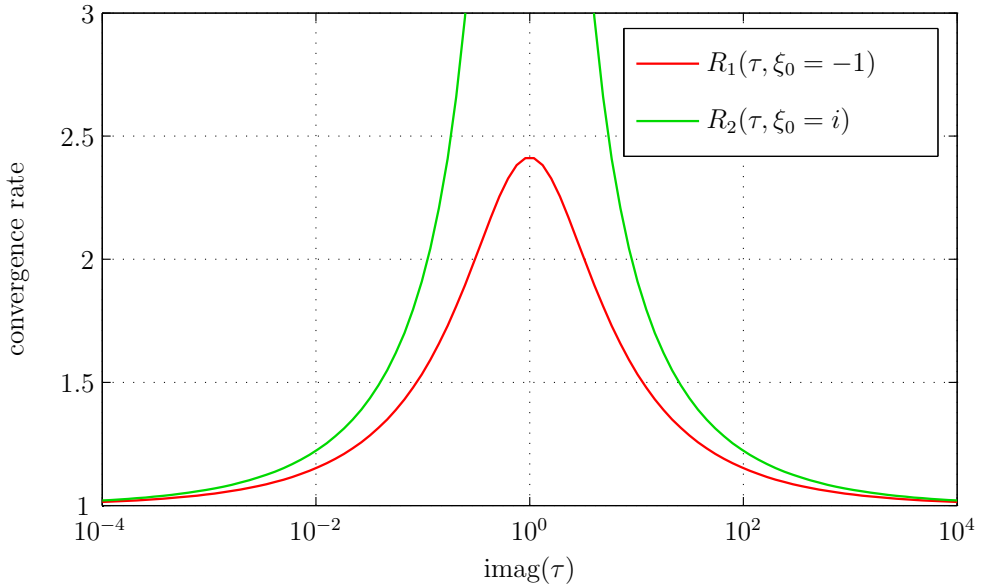


Figure 7.6: Graph of the functions $R_1(\tau, \xi_0)$ and $R_2(\tau, \xi_0)$, which describe the asymptotic convergence rates for approximating the resolvent function f^τ using a single repeated real or imaginary pole ξ_0, respectively.

is the rational function of type $(m-1, m-1)$ with the same poles as s_m that interpolates the resolvent function (7.12) at the zeros of s_m. The absolute error is explicitly given by

$$f^\tau(z) - r_m^\tau(z) = \frac{s_m(z)}{s_m(\tau)} \frac{1}{z-\tau}, \tag{7.17}$$

and the relative error takes the even simpler form

$$\big[f^\tau(z) - r_m^\tau(z)\big] / f^\tau(z) = \frac{s_m(z)}{s_m(\tau)}. \tag{7.18}$$

(The same representation for the error can be derived with the help of so-called *skeleton approximations* of f^τ, see [Tyr96, KDZ09].)

If the set of allowed poles Ξ coincides with the parameter set T, the minimization of (7.18) for all $z \in \Sigma$ is precisely the Zolotarev problem we considered in Theorem 7.5, and hence the zeros and poles of s_m should be asymptotically distributed according to the equilibrium measure of the condenser (Σ, Ξ) (e.g., one could use generalized Leja points). In [KDZ09] the authors explicitly determine optimal rational functions s_m^* for the sets $\Sigma = [0, +\infty)$ and $\Xi = T' = [-\tau_{\min}, -\tau_{\max}] \cup [\tau_{\min}, \tau_{\max}]$, the latter being purely imaginary and symmetric

with respect to the real axis. They also give the expression for the convergence rate as

$$R' = \exp\left(\frac{\pi}{2}\frac{K(\kappa)}{K(1-\kappa)}\right), \quad \text{where} \quad \kappa = \frac{\tau_{\min}}{\tau_{\max}}$$

and $K(\kappa) = \int_0^1 [(1-t^2)(1-\kappa t^2)]^{-1/2}\,\mathrm{d}t$. The proof in [KDZ09] relates the zeros and poles of s_m^* to those of the Zolotarev approximation for the square root on the positive interval $[|\tau_{\min}|, |\tau_{\max}|]$. This technique relies strongly on the symmetries of T' and the Green's function $g_\Omega(z,\zeta)$ for $\Omega = \mathbb{C}\backslash[0,+\infty)$ with pole at ζ. However, since $f^{-\tau}(z) = \overline{f^\tau}(z)$ for real arguments z, it actually suffices to consider the half condenser plate $\Xi = T = [\tau_{\min}, \tau_{\max}]$, which should give rise to an improved convergence rate $R > R'$. Unfortunately, the loss of symmetry makes it more complicated to construct the conformal mapping of $\mathbb{C} \setminus (\Sigma \cup \Xi)$ onto the annulus $\mathbb{A}_R$ and to obtain a formula for R. It should be possible to construct such a mapping via the Schwarz–Christoffel formula for doubly connected regions (see [Hen93, §17.5], [Däp88]).

Instead of following this approach further we propose a simple and constructive method for generating a sequence of cyclically repeated asymptotically "good" poles. This method will also have the advantage that Ξ needs not necessarily coincide with T. Neither is it required that Σ be an interval.

7.5.3 Cyclically Repeated Poles

As before we consider $\Sigma = [0, +\infty]$ and $T = [\tau_{\min}, \tau_{\max}]$ positive imaginary. Recall the function $R_1(\tau, \xi_0)$ [or $R_2(\tau, \xi_0)$] in Remark 7.11, which gives the asymptotic convergence rate one can expect from optimal rational approximation of the function $f^\tau(z) = (z-\tau)^{-1}$ with all poles at the point $\xi_0 < 0$ [or $i\xi_0 < 0$]. In other words, one iteration of a near-optimal extraction (such as Rayleigh–Ritz) from a rational Krylov space with all poles at ξ_0 reduces the error by $R_1(\tau, \xi_0)^{-1}$ [or $R_2(\tau, \xi_0)^{-1}$].

Consider now p poles $\xi_1, \ldots, \xi_p$ being all contained in the *real* interval $\Xi = iT$. The product form (7.17) of the error allows us to conclude that the Rayleigh–Ritz method with these poles repeated cyclically converges (at least) at the rate

$$R(\tau) = [R_1(\tau, \xi_1) \cdots R_1(\tau, \xi_p)]^{1/p},$$

depending on the parameter $\tau \in T$. Obviously, it is desirable to make the minimum of $R(\tau)$ as large as possible on T (the factors $R_1(\tau, \cdot)$ are continuous on the compact set T and hence attain a minimum, cf. Figure 7.6). A simple method for finding a suitable placement of real poles is given by successive maximization of the worst-case convergence rate.

(a) Set $j := 1$ and initialize a function $\widetilde{R}(\tau) \equiv 1$ for all $\tau \in T$.

(b) Choose ξ_j such that

$$\min_{\tau \in T} \widetilde{R}(\tau) R_1(\tau, \xi_j) = \max_{\xi \in \Xi} \min_{\tau \in T} \widetilde{R}(\tau) R_1(\tau, \xi)$$

(c) Update $\widetilde{R}(\tau) := \widetilde{R}(\tau) R_1(\tau, \xi_j)$ for all $\tau \in T$.

(d) Set $j := j + 1$.

(e) If $j \leq p$ go to (b).

Note that step (b) is computationally difficult to handle, except if T and Ξ are discrete sets with only a few elements. For we know that $R_1(\tau, \xi)$ attains its maximum value on T at the point $\tau = \xi/i$ (cf. Example 7.9), it is reasonable to replace step (b) by

(b') Choose ξ_j such that $\widetilde{R}(\xi_j/i) = \min_{\tau \in T} \widetilde{R}(\tau)$.

The resulting method is very easily implemented, e.g., in the chebfun system. We remark that this method works with slight modifications if $R_1(\tau, \xi)$ is replaced by $R_2(\tau, \xi)$. In fact, it works if only a formula for the asymptotic error reduction as a function of $\tau \in T$ and $\xi \in \Xi$ is available.

Example 7.12. We compare the "approximation power" of a rational Krylov space with real poles computed by the proposed method and a rational Krylov space with purely imaginary poles on $\Xi = T$ (generalized Leja points). To this end we consider a symmetric matrix A with equispaced eigenvalues in $[0, 10^4] \subset \Sigma = [0, +\infty]$ and a parameter interval $T = i[1, 10^3]$. We approximate the resolvent function f^τ for 11 logspaced parameters $\tau \in T$ using the Rayleigh–Ritz method. In Figure 7.7 we show the resulting convergence curves and the asymptotic convergence rates using real and imaginary poles, respectively. As expected, the method with the imaginary poles converges at a higher rate as the method with the real poles. Note that with imaginary poles the error may drop drastically at the jth iteration, which happens if a pole ξ_j hits one of the parameters τ exactly (e.g., the extremal ones). However, the faster convergence of this method comes at the price

of solving linear systems with a complex shift (i.e., complex arithmetic) and one may ask whether this pays off. We try to answer this question in Figure 7.8, where we show the (estimated) asymptotic convergence rates $R_1(c)$ and $R_2(c)$ of the methods with real and imaginary poles for parameter sets $T = [i, ic]$ with varying ratio c, respectively. It is surprising how fast the logarithmic quotient of both rates decays as c gets larger. For example, if one iteration with an imaginary pole is 4 times as expensive as with a real pole, it is advised to use real poles already if $c \gtrapprox 1.41$ (in practical applications the ratio c is usually larger than, say, 10^3).

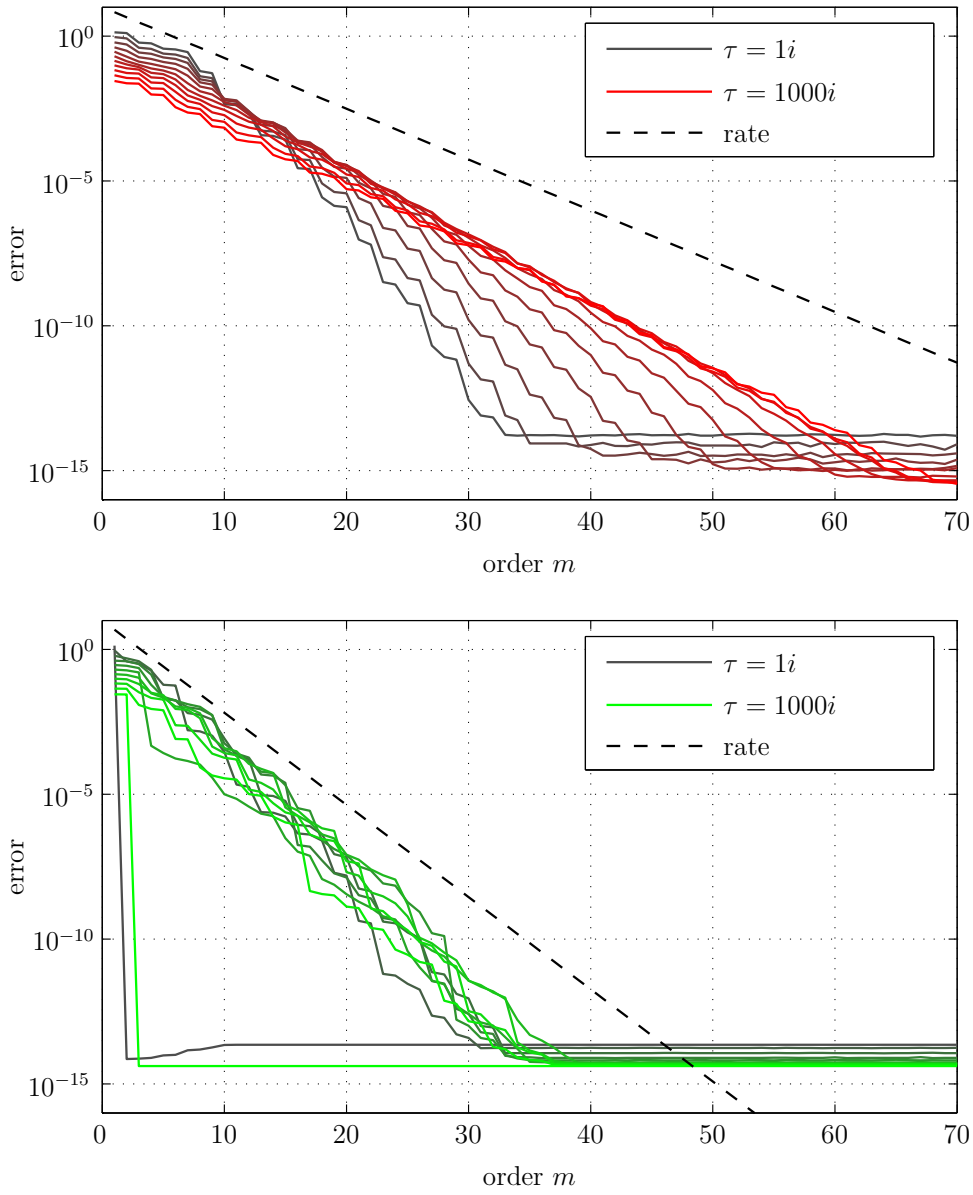


Figure 7.7: Error curves (solid) and asymptotic convergence rates (dashed) of Rayleigh–Ritz approximations for the resolvent function extracted from a rational Krylov space with real poles (top) and imaginary poles (bottom).

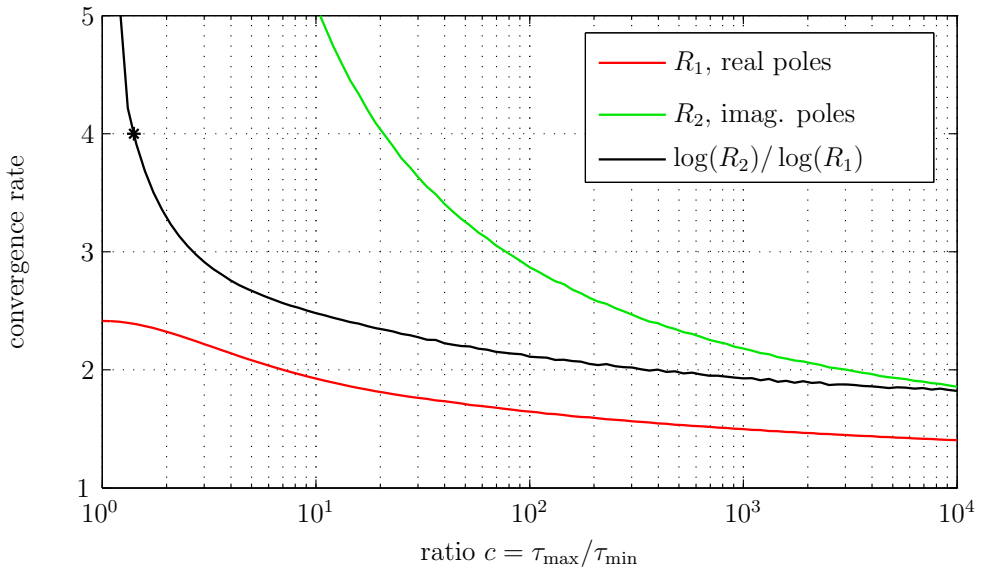


Figure 7.8: Estimated asymptotic convergence rates on the parameter set $T = [\tau_{\min}, \tau_{\max}]$ as a function of $c = \tau_{\max}/\tau_{\min}$ using real poles (R_1, red) compared to imaginary poles (R_2, green). The logarithmic quotient (black) indicates how many iterations are required by Rayleigh–Ritz extraction from a rational Krylov space with optimal real poles to achieve the same error reduction as one iteration with optimal imaginary poles.

7.6 The Exponential Function

In this section we approximate the exponential function

$$f^\tau(z) = \exp(\tau z)$$

on a closed set Σ by rational functions $r_m^\tau = p_{m-1}^\tau / q_{m-1}$ with poles in a closed set Ξ. As before, the parameters τ are collected in a parameter set T. This parameter set is often real as τ usually corresponds to time.

The exponential function represents the exact solution $\boldsymbol{u}(\tau) = \exp(\tau A)\boldsymbol{u}_0$ of the most fundamental dynamical system

$$\boldsymbol{u}'(\tau) = A\boldsymbol{u}(\tau), \quad \boldsymbol{u}(0) = \boldsymbol{u}_0,$$

hence is of high relevance, e.g., for exponential integrators [HLS98, CM02, KT05, MW05, ST07b] or in Markov chain analysis [PS93, Saa95, SS99], and has numerous applications

in nuclear magnetic resonance spectroscopy [NH95], quantum physics [NW83], and geophysics [DK94], to name just a few. In all these applications, polynomial and rational Krylov order reduction techniques have proved to be very efficient, especially in large-scale computations.

7.6.1 Real Pole Approximations

Due to their computational advantages, rational approximations to the exponential function on $(-\infty, 0]$ with real poles in $\Xi = \mathbb{R}_+$ have been studied extensively in the literature, see [SSV76, Lau77, NW77, Ser92]. In particular, it was shown in [Bor83] that the best uniform approximation in $\mathcal{R}^{\Xi}_{m,m}$ to $f^1(z) = \exp(z)$ on $(-\infty, 0]$ has a pole of order m. A result due to [And81] states that this pole behaves asymptotically like $m/\sqrt{2}$.

The Parameter-Free Case. Let us consider the approximation of f^1 on an interval $\Sigma = [a, b]$, $-\infty < a < b \leq 0$ by a rational function with a single repeated pole $\xi > 0$. A possible approach for obtaining the linear convergence rate is to again reduce the problem to polynomial approximation. As proposed in Section 7.4, set $\hat{z} = (z-\xi)^{-1}$ and approximate $\hat{f}(\hat{z}) := \exp(\hat{z}^{-1} + \xi)$ by a polynomial on $\widehat{\Sigma} = [(b-\xi)^{-1}, (a-\xi)^{-1}]$. As $\hat{f}$ has its only singularity at $\hat{z} = 0$, the asymptotic convergence rate should be the same as for the CG method applied to a symmetric matrix with positive spectral interval $-\widehat{\Sigma}$, i.e.,

$$R = \exp(g_{\mathbb{C}\setminus\widehat{\Sigma}}(0)) \approx \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \quad \text{with } \kappa = (a-\xi)/(b-\xi).$$

Note that the rate R increases as ξ goes to infinity. This corresponds to the fact that polynomial best uniform approximation of the exponential function on Σ converges superlinearly, not just linearly like polynomial approximation of $\hat{f}$ on $\widehat{\Sigma}$. However, we are not primarily interested in the asymptotic rate but in the error of polynomials of finite degree. The asymptotic convergence theory of polynomial approximation does not reveal that fast linear convergence in early iterations can be much better than superlinear convergence for late iterations. Another limitation is that this asymptotic theory is not applicable when $a = -\infty$, in which case $\hat{f}$ is not analytic in any ellipse containing $\widehat{\Sigma}$. Let us therefore employ another approach in the following, covering more generally the approximation of the *parameter-dependent* function $f^\tau(z) = \exp(\tau z)$.

The Parameter-Dependent Case. Let the configuration $\Sigma = [-\infty, 0]$, $T = [\tau_{\min}, \tau_{\max}] \subset \mathbb{R}_+$, and $\Xi = \{\xi_1, \ldots, \xi_p\} \subset \mathbb{R}_+$ be given. How should the poles in Ξ be chosen such that we can guarantee $\|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\| \leq 2\varepsilon$ for all $\tau \in T$, where $\boldsymbol{f}_m^\tau$ is the Rayleigh–Ritz approximation from a rational Krylov space $\mathcal{Q}_m$ with poles in Ξ? The situation is more complicated than it was for the resolvent function since a simple error formula of type (7.17) does not exist. However, with the help of the following corollary one can still justify the fact that cyclically repeated poles yield at least as good approximations as the single poles could achieve taken separately.

Corollary 7.13. *Assume that the rational Krylov space $\mathcal{Q}_m$ contains the pole ξ at least $n-1$ times, that is,*

$$\mathcal{S}_n := \operatorname{span}\{\boldsymbol{b}, (A-\xi I)^{-1}\boldsymbol{b}, \ldots, (A-\xi I)^{-n+1}\boldsymbol{b}\} \subseteq \mathcal{Q}_m.$$

Then the Rayleigh–Ritz approximation $\boldsymbol{f}_m^\tau$ for $f^\tau(A)\boldsymbol{b}$ from $\mathcal{Q}_m$ satisfies

$$\|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\| \leq 2C \min_{p_{n-1}\in\mathcal{P}_{n-1}} \|f^\tau(z) - p_{n-1}(z)/(z-\xi)^{n-1}\|_\Sigma,$$

with $\Sigma \supseteq \mathbb{W}(A)$ and a constant $C \leq 11.08$.

If A is self-adjoint the result holds with $C = 1$.

Proof. This follows from Theorem 4.10 and the fact that $\mathcal{S}_n \subseteq \mathcal{Q}_m$. □

The following approach is a straightforward generalization of a method used in [EH06] (see also [PS08] for related work). Adhering to the notation introduced in Section 7.4, we have

$$\begin{aligned} \eta_n(\tau,\xi) &:= \min_{p_{n-1}\in\mathcal{P}_{n-1}} \|f^\tau(z) - p_{n-1}(z)/(z-\xi)^{n-1}\|_\Sigma \\ &= \min_{\hat{p}_{n-1}\in\mathcal{P}_{n-1}} \|\hat{f}^\tau(\hat{z}) - \hat{p}_{n-1}(\hat{z})\|_{\widehat{\Sigma}}, \end{aligned} \tag{7.19}$$

where $\hat{z} = (z-\xi)^{-1}$, $\widehat{\Sigma} = [-\xi^{-1}, 0]$, and $\hat{f}^\tau(\hat{z}) = e^{\tau(\hat{z}^{-1}+\xi)}$. Obviously, there holds $\eta_n(\tau,\xi) = \eta_n(c\tau, \xi/c)$ for all $c > 0$. This means that if we have a partition $T = T_1 \cup \cdots \cup T_p$ into p intervals of the form $T_j = \tau_{\min}[c^{j-1}, c^j]$ and we find a pole ξ_1 such that $\eta_n(\tau,\xi_1) \leq \varepsilon$ for all $\tau \in T_1$, then we know that $\eta_n(\tau, \xi_1/c^{j-1}) \leq \varepsilon$ for all $\tau \in T_j$.

The above suggests the use of the poles

$$\xi_j := \frac{\xi_1}{c^{j-1}}, \quad j = 1, \dots, p,$$

since by Corollary 7.13 we then have

$$\|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\| \leq 2C\varepsilon, \quad \text{for all } \tau \in T,$$

where $m = p(n-1)+1$ and $\boldsymbol{f}_m$ is the Rayleigh–Ritz approximation from a Krylov space $\mathcal{Q}_m$ with each pole ξ_j repeated cyclically for $n-1$ times.

Example 7.14. Let $T = [10^{-3}, 1]$ and $p = 3$, so that we have $c = 10$ and $T = [10^{-3}, 10^{-2}] \cup [10^{-2}, 10^{-1}] \cup [10^{-1}, 1]$. We aim for an approximation error $\|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\|$ smaller than $2\varepsilon = 2 \cdot 10^{-7}$. One verifies with the help of Figure 7.9 (top) that $\eta_{20}(\tau, \xi_0 = 1) \leq \varepsilon$ for all $\tau \in T_0 = [3.25, 32.7]$. This plot has been generated by running the Remez algorithm for various τ (we used the chebfun implementation of the Remez algorithm given in [PT09]). Rescaling T_0 and ξ_0 yields

$$\xi_1 = 3.25 \cdot 1000, \quad \xi_2 = 3.25 \cdot 100, \quad \xi_3 = 3.25 \cdot 10$$

as the desired poles.

In Figure 7.9 (bottom) we show the error curves of Rayleigh–Ritz approximations for 11 logspaced parameters $\tau \in [10^{-3}, 1]$, with a symmetric matrix A with equispaced eigenvalues in $[-10^5, 0]$. As expected, the errors do not exceed the level $2 \cdot 10^{-7}$ for all iterations of order larger than $m = p(n-1)+1 = 3(20-1)+1 = 58$, which is indicated by the green bar.

Remark 7.15. The above approach is constructive and guarantees that in a certain iteration m the error $\|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\|$ is smaller than a prescribed tolerance for all $\tau \in T$. In particular, if $\Sigma = [-\infty, 0]$ is taken unbounded as in our Example 7.14, this error tolerance will be respected for every symmetric negative semi-definite operator A. On the other hand, we make use of the somewhat "pessimistic" assumption that a pole ξ_j contributes only to those approximations $\boldsymbol{f}_m^\tau$ for which $\tau \in T_j$. For very large parameter intervals T this inevitably yields rational approximations of order much higher than actually required. In this case it is advisable to use another approach for computing the poles, which we describe in the following.

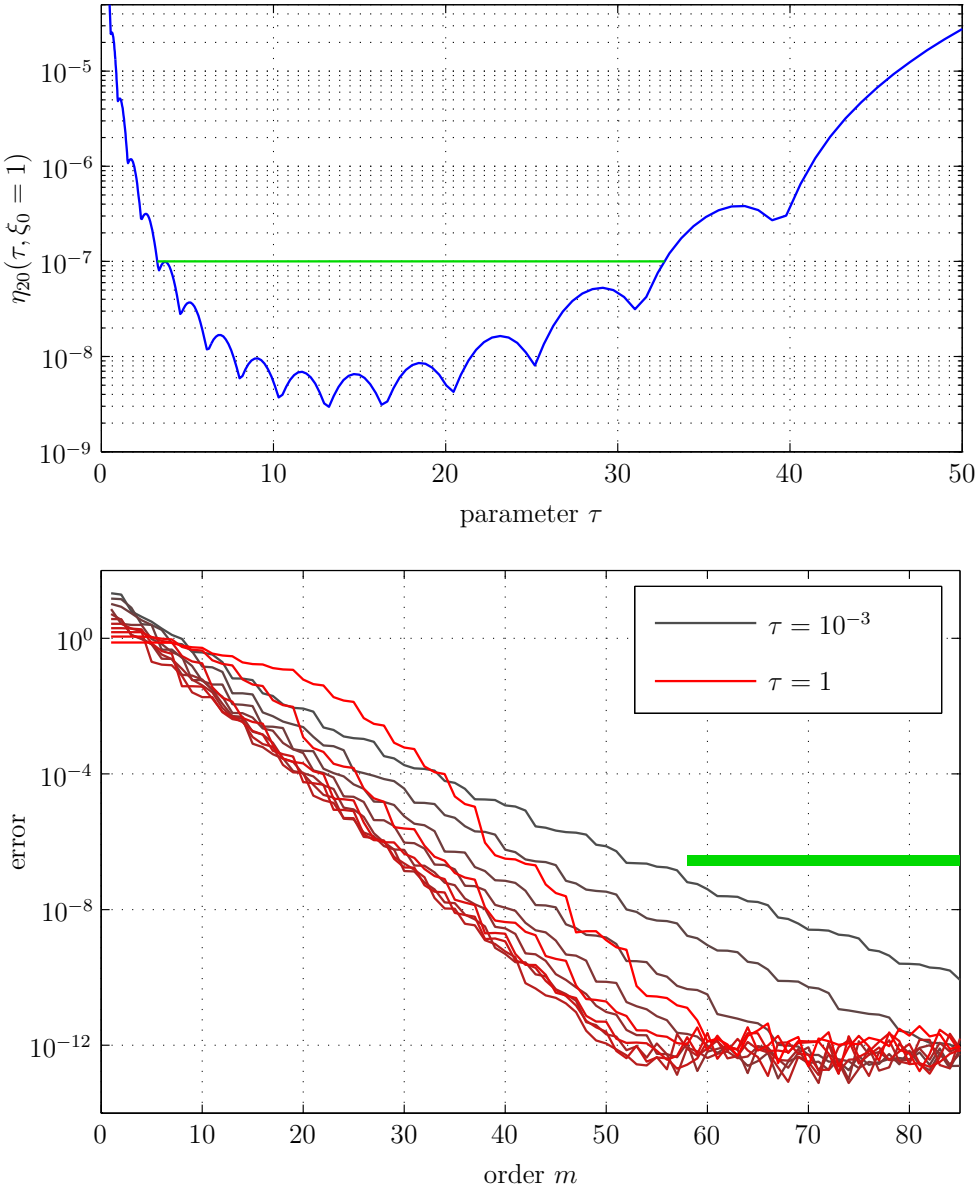


Figure 7.9: The function $\eta_{20}(\tau, \xi_0 = 1)$ (top) defined in (7.19) and convergence curves of Rayleigh–Ritz approximations for $f^\tau(A)\boldsymbol{b} = \exp(\tau A)\boldsymbol{b}$ with 11 logspaced parameters $\tau \in [10^{-3}, 1]$ (bottom). The poles of the rational Krylov space are repeated cyclically. By construction, the error curves stay below the green bar.

7expcycl

7.6.2 Connection to a Zolotarev Problem

We sketch a novel approach for constructing rational approximations to f^τ on a *bounded* interval $\Sigma = [a, b]$, $a < b < 0$, which has been presented recently by Druskin, Knizhnerman & Zaslavsky [DKZ09]. These rational functions have real poles and converge at a fixed rate R to f^τ for arbitrary parameters τ. The rate R depends on the ratio b/a, and hence this approach should be preferred to the one described in the previous section if b/a is moderate and T is a large interval.

The key idea is to exploit the inverse Fourier representation of the exponential function, namely

$$f^\tau(z) = \exp(\tau z) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{\exp(\tau\zeta)}{\zeta - z} \,\mathrm{d}\zeta, \quad z < 0,\ \tau \geq 0$$

(which could also be called *Bromwich integral* [WT07]). Using this integral representation for $f^\tau(A)\boldsymbol{b}$ and the Rayleigh–Ritz approximation $\boldsymbol{f}_m^\tau = V_m f^\tau(A_m) V_m^\dagger \boldsymbol{b}$, one obtains

$$\begin{aligned} f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau &= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \exp(\tau\zeta) \left[(\zeta I - A)^{-1}\boldsymbol{b} - V_m(\zeta I_m - A_m)^{-1} V_m^\dagger \boldsymbol{b} \right] \mathrm{d}\zeta \\ &= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \exp(\tau\zeta) \left[r^\zeta(A)\boldsymbol{b} - r_m^\zeta(A)\boldsymbol{b} \right] \mathrm{d}\zeta, \end{aligned}$$

where $r_m^\zeta(z) = p_{m-1}^\zeta(z)/q_{m-1}(z)$ interpolates the resolvent function $r^\zeta(z) := (\zeta - z)^{-1}$ at the nodes $\Lambda(A_m)$. The aim is to make the error $r^\zeta - r_m^\zeta$ as small as possible on Σ for all "parameters" $\zeta \in i\mathbb{R}$. As described in Section 7.5.2, this can be done by solving a Zolotarev problem on the condenser (Σ, Ξ) with plates Σ and $\Xi = \{z : \Re(z) \geq 0\}$. Again with the help of elliptic functions one can show that $\mathbb{C} \setminus (\Sigma \cup \Xi)$ is conformally equivalent to the annulus $\mathbb{A}_R$ with

$$R = \exp\left(\frac{\pi}{4} \frac{K(1-\kappa)}{K(\kappa)} \right), \quad \text{where } \kappa = \left(\frac{1 - \sqrt{a/b}}{1 + \sqrt{a/b}} \right)^4, \tag{7.20}$$

$K(\kappa) = \int_0^1 [(1-t^2)(1-\kappa t^2)]^{-1/2} \,\mathrm{d}t$. The near-optimality of Rayleigh–Ritz approximations then allows us to conclude that

$$\limsup_{m\to\infty} \|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\|^{1/m} \leq R^{-1},$$

if the poles of the rational Krylov space are distributed on $\partial\Xi = i\mathbb{R}$ according to the

equilibrium measure of the condenser (Σ, Ξ).

It has been proven in [LT00] that for this special geometry of a real and an imaginary interval symmetric to the real axis there is a sequence of rational functions $s_m \in \mathcal{R}^{-\Sigma}_{m,m-1}$ with poles in $-\Sigma = [-b, -a] \in \mathbb{R}_+$ and zeros in Σ that satisfy

$$\limsup_{m\to\infty} \left(\frac{\sup_{z\in\Sigma} |s_m(z)|}{\inf_{z\in i\mathbb{R}} |s_m(z)|} \right)^{1/m} = R^{-1},$$

where R is the same as in (7.20). It is important to underline that the poles of s_m are real, and hence computationally convenient for building a rational Krylov space.

Example 7.16. In Figure 7.10 we show the error curves of Rayleigh–Ritz approximations for $f^\tau(A)\boldsymbol{b} = \exp(\tau A)\boldsymbol{b}$ for 11 logspaced parameters $\tau \in T = [10^{-3}, 1]$, where A is a symmetric negative definite matrix with equispaced eigenvalues in $\Sigma = [-10^5, -1]$. We compute generalized Leja points on the condenser $(\Sigma, -\Sigma)$ and use the points on the plate $-\Sigma$ as poles for the rational Krylov space. We also show the asymptotic convergence rate (7.20).

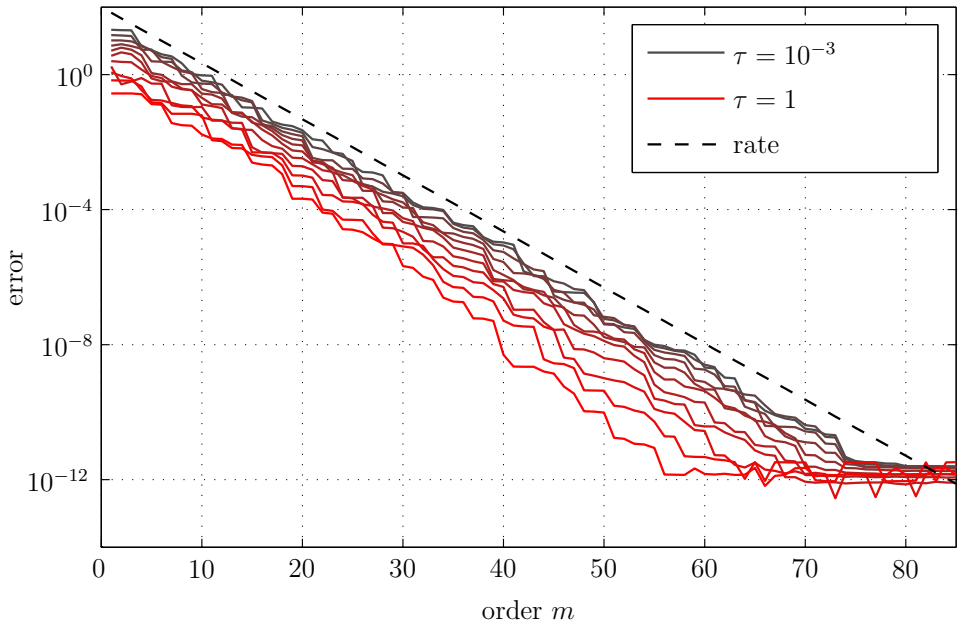


Figure 7.10: Convergence curves and asymptotic convergence rate of Rayleigh–Ritz approximations for $f^\tau(A)\boldsymbol{b} = \exp(\tau A)\boldsymbol{b}$ with 11 logspaced parameters $\tau \in T = [10^{-3}, 1]$. The poles of the rational Krylov space lie on the positive real axis and are optimal for a certain Zolotarev problem.


7expzol

7.7 Other Functions, Other Techniques

Rational approximation theory is a very rich field and it is impossible to give a complete treatment here. Usually, every function f requires a special investigation depending on the configuration of Σ, T and Ξ. This makes the choice of parameters in rational Krylov methods for $f(A)\boldsymbol{b}$ challenging, and also interesting. We have already discussed important issues, such as the choice of poles and nodes for the PAIN method, or the real pole selection when approximating the resolvent and exponential function. In this brief section we give some references to other approximation techniques that may be applicable, without being exhaustive.

First of all, we repeat that the poles of a rational best approximation to f on Σ are good candidates for poles of a rational Krylov space when using the near-optimal Rayleigh–Ritz approximations. In a few special cases rational best approximations r_m^* are explicitly known, the most prominent example being the best relative approximation to $f(z) = z^{1/2}$ on a positive real interval and the related best uniform approximation to $\operatorname{sgn}(z)$ on two disjoint real intervals constructed by Zolotarev, see [Zol77, Tod84]. For work relating matrix functions and Zolotarev's approximations we refer to [EFL$^+$02, Ken04]. Seven possible ways to obtain rational approximations for the function $f(z) = (1 - z^2)^{1/2}$ on $\Sigma = [-1, 1]$, which is connected with the so-called one-way wave equation, are investigated in [HT88]. An error estimate for best uniform rational approximation of z^α on $\Sigma = [0, 1]$ is given in [Sta03].

Another problem that has attracted a lot of attention is the rational approximation of the exponential function $f(z) = \exp(z)$, typically on the unit disk $\mathbb{D}$ or on the negative real axis $(-\infty, 0]$, see [CMV69]. Recently the exponential function and the related φ-functions have gained special interest in connection with exponential integrators. The practical computation of near-best approximations using the Carathéodory–Fejér method and contour integrals is considered in [TWS06, ST07a, ST07b, Sch07, HHT08]. The Carathéodory–Fejér method for rational approximation on the unit disk is based on a singular value decomposition of a Hankel matrix set up with a few Taylor coefficients of the function f, and it provides rational approximations closer to best than can practicably be achieved by other means [Tre81, Tre83]. In [TG83] it is shown that in conjunction with a conformal transplantation this method is efficient for real approximation, too. On the other hand,

for parameter-dependent functions f^τ rational approximations obtained by quadrature formulas on contour integrals may sometimes perform better than Carathéodory–Fejér approximations [TWS06].

The approximation of trigonometric functions of symmetric positive definite operators by the shift-and-invert method is considered in [GH08]. Clearly, no rational function can be a good approximation for an oscillating function on an infinite interval, so the introduction of an artificial decay factor becomes necessary. The connection of the shift-and-invert method to polynomial approximation in conjunction with Jackson's approximation theorems [Che98, Sec. 4.6] allows one to give bounds for the approximation error. Padé approximations of trigonometric functions were considered in [Col89], and approximations with a repeated pole in [DS80].

We demonstrated in Section 4.2.2 how Crouzeix's theorem can be used to bound the error of the Rayleigh–Ritz approximation $\boldsymbol{f}_m$ via the uniform error of the rational best uniform approximation $r_m^* \in \mathcal{P}_{m-1}/q_{m-1}$ to f on $\Sigma \supseteq \mathbb{W}(A)$. This yields an a-priori error estimate if one is able to bound the error $\|f - r_m^*\|_\Sigma$. For Markov functions f, explicit lower and upper bounds for $\|f - r_m^*\|_\Sigma$ involving finite Blaschke products are derived in [BR09] by making use of the Faber transform. These bounds can also be used for pole optimization. In general, the Faber transform is a powerful tool for constructing rational approximations to analytic functions on simply connected sets Σ [Ell83]. We remark that the expansion of a function in Faber polynomials can be used directly to obtain efficient polynomial Krylov methods with asymptotically optimal error reduction [HPKS99, MN01a, Nov03, BCV03], a particular interesting special case on intervals being expansions in Chebyshev polynomials [DK89, Sch90, CRZ99, BV00].

8 Rational Ritz Values

This chapter is adopted in parts from [BGV10].

The rational Ritz values, which are the eigenvalues of the Rayleigh quotient $A_m = V_m^\dagger A V_m$, have played a prominent role so far, for example as interpolation nodes in Rayleigh–Ritz approximations. It is known that the Ritz values tend to approximate eigenvalues of A in proximity of the poles of the rational Krylov space. We will give a theoretical explanation of this behavior when A is Hermitian. In Section 8.1 we investigate how the distance between Ritz values and eigenvalues of A can be bounded via a polynomial extremal problem. In the following sections we describe in an asymptotic sense *which* eigenvalues of A are approximated by rational Ritz values and *how fast* this approximation takes place. This description gives insight into the rational Arnoldi algorithm used to compute eigenvalues of Hermitian matrices, but it can also explain superlinear convergence effects observed with Rayleigh–Ritz approximations for $f(A)\boldsymbol{b}$. We decided to use a more intuitive approach to this asymptotic theory, since many details are involved and a rigorous derivation can be found in [BGV10].

8.1 A Polynomial Extremal Problem

In what follows we consider a Hermitian matrix $A \in \mathbb{C}^{N\times N}$ with distinct eigenvalues $\lambda_1 < \cdots < \lambda_N$. We assume that $A = UDU^*$ is the spectral decomposition of A with normalized eigenvectors $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N]$ and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$. The normalized eigencomponents of $\boldsymbol{b}$ in the basis U are denoted by $w(\lambda_j)$, i.e.,

$$w(\lambda_j) = |\langle \boldsymbol{u}_j, \boldsymbol{b}/\|\boldsymbol{b}\|\rangle| \in [0,1], \quad j = 1, \ldots, N. \tag{8.1}$$

By Θ we denote the set of mth rational Ritz values $\theta_1 < \cdots < \theta_m$ associated with the rational Krylov space $\mathcal{Q}_m = q_{m-1}(A)^{-1}\mathcal{K}_m(A, \boldsymbol{b})$, and χ_m is the associated nodal polynomial $\chi_m(x) = (x-\theta_1)\cdots(x-\theta_m)$. By Ξ we now denote the *multiset* of poles $\xi_1, \ldots, \xi_{m-1}$, which are the zeros of q_{m-1}.

It is often observed in practice that rational Ritz values tend to approximate some of A's eigenvalues very quickly. The natural question arises, which eigenvalues are approximated by rational Ritz values of order m? In other words, can we give bounds for the distance of an eigenvalue λ_k to the set Θ? It is clear that this distance cannot be small for all eigenvalues λ_k since there are fewer Ritz values than eigenvalues. In several textbooks [GV96, PPV95, Saa92b, TB97] one can find bounds for polynomial Ritz values and these results are classical now. Many of them are derived by exploiting the relationship between polynomials and Krylov spaces, where an important ingredient for estimating the distance of an eigenvalue to the set of Ritz values is a link to some polynomial extremal problem. Typically, such procedures are used to handle extremal eigenvalues or outliers, but this approach is also useful for eigenvalues in other parts of the spectrum [Bec00a]. The following result, also given in [BGV10, Lemma 2.3], is an extension of [Bec00a, Lemma 2.2] to rational Ritz values. Here we give a different proof.

Lemma 8.1. *If $\lambda_k \leq \theta_1$ then*

$$\theta_1 - \lambda_k = \min \left\{ \frac{\sum_{j=1, j\neq k}^{N} \frac{w(\lambda_j)^2}{q_{m-1}(\lambda_j)^2} (\lambda_j - \theta_1)\, s_{m-1}(\lambda_j)^2}{\frac{w(\lambda_k)^2}{q_{m-1}(\lambda_k)^2} s_{m-1}(\lambda_k)^2} \;:\; s_{m-1} \in \mathcal{P}_{m-1},\ s_{m-1}(\lambda_k) \neq 0 \right\}.$$

The minimum is attained for $s_{m-1}(x) = \chi_m(x)/(x-\theta_1)$.

If $\lambda_k \in [\theta_{\kappa-1}, \theta_\kappa]$ then

$$\begin{aligned} & (\lambda_k - \theta_{\kappa-1})\,(\theta_\kappa - \lambda_k) \\ = \quad & \min \left\{ \frac{\sum_{j=1, j\neq k}^{N} \frac{w(\lambda_j)^2}{q_{m-1}(\lambda_j)^2} (\lambda_j - \theta_{\kappa-1})\,(\lambda_j - \theta_\kappa)\, s_{m-2}(\lambda_j)^2}{\frac{w(\lambda_k)^2}{q_{m-1}(\lambda_k)^2} s_{m-2}(\lambda_k)^2} \;:\; s_{m-2} \in \mathcal{P}_{m-2},\ s_{m-2}(\lambda_k) \neq 0 \right\}. \end{aligned}$$

The minimum is attained for $s_{m-2}(x) = \chi_m(x)/\left((x-\theta_{\kappa-1})(x-\theta_\kappa)\right)$.

Proof. We only prove the second part $\lambda_k \in [\theta_{\kappa-1}, \theta_\kappa]$; the proof for the other part is similar. Without loss of generality we assume that V_m is an orthonormal basis of $\mathcal{Q}_m$ and

$A_m = V_m^* A V_m$. We set $\boldsymbol{q} := q_{m-1}(A)^{-1}\boldsymbol{b}$ and first show that

$$\alpha = \langle s_{m-2}(A)\boldsymbol{q}, (A-\theta_{\kappa-1}I)(A-\theta_\kappa I)s_{m-2}(A)\boldsymbol{q}\rangle \tag{8.2}$$

is a nonnegative real number for all $s_{m-2} \in \mathcal{P}_{m-2}$. Since $s_{m-2}(A)\boldsymbol{q} \in \mathcal{Q}_m$, we can use the exactness of Rayleigh–Ritz approximations (cf. Lemma 4.6) to obtain

$$\begin{aligned}
\alpha &= \langle s_{m-2}(A)\boldsymbol{q}, P_m(A-\theta_{\kappa-1}I)(A-\theta_\kappa I)s_{m-2}(A)\boldsymbol{q}\rangle \\
&= \langle s_{m-2}(A_m)V_m^*\boldsymbol{b}, (A_m-\theta_{\kappa-1}I_m)(A_m-\theta_\kappa I_m)s_{m-2}(A_m)V_m^*\boldsymbol{b}\rangle \\
&= \langle \boldsymbol{x}, (A_m-\theta_{\kappa-1}I_m)(A_m-\theta_\kappa I_m)\boldsymbol{x}\rangle,
\end{aligned}$$

where $P_m = V_m V_m^*$ is the orthogonal projector onto $\mathcal{Q}_m$ and $\boldsymbol{x} := s_{m-2}(A_m)V_m^*\boldsymbol{b}$. Let $A_m = X_m D_m X_m^*$ be a spectral decomposition of A_m with $X_m^* X_m = I_m$ and $D_m = \operatorname{diag}(\theta_1, \ldots, \theta_m)$, then

$$\alpha = \sum_{j=1}^{m} (\theta_j - \theta_{\kappa-1})(\theta_j - \theta_\kappa)|\boldsymbol{e}_j X_m^* \boldsymbol{x}|^2,$$

which is obviously a nonnegative real number. Note that, by Lemma 4.5 (b), we get $\alpha = 0$ in (8.2) if s_{m-2} is chosen such that $(x-\theta_{\kappa-1})(x-\theta_\kappa)s_{m-2}(x) = \chi_m(x)$. By replacing $A = UDU^*$ in (8.2) we find

$$\alpha = \sum_{j=1}^{N} \frac{w(\lambda_j)^2}{q_{m-1}(\lambda_j)^2}(\lambda_j - \theta_{\kappa-1})(\lambda_j - \theta_\kappa)s_{m-2}(\lambda_j)^2,$$

from which the assertion of the lemma follows by separating the term with $j = k$. □

To give a better understanding of the potential impact of Lemma 8.1, let us have a closer look at the first part for polynomial Ritz values (i.e., $q_m \equiv 1$). Since all Ritz values lie in the open interval (λ_1, λ_N), we may choose $k = 1$ and obtain for $\operatorname{dist}(\lambda_1, \Theta) = \theta_1 - \lambda_1$ the upper bound

$$\operatorname{dist}(\lambda_1, \Theta) \le |\lambda_N - \lambda_1| \frac{\max_{j=2,\ldots,N} |s_{m-1}(\lambda_j)|^2}{|s_{m-1}(\lambda_1)|^2} \sum_{j=2}^{N} \frac{w(\lambda_j)^2}{w(\lambda_1)^2}$$

for any polynomial $s_{m-1} \in \mathcal{P}_{m-1}$ with $s_{m-1}(\lambda_1) \neq 0$. More explicit upper bounds are

obtained by choosing s_{m-1} to take the value 1 at λ_1 and be small on the convex hull of all other eigenvalues, leading to the well-known Kaniel–Page–Saad estimate for extremal eigenvalues [GV96, PPV95, Saa92b, TB97]. This construction is similar to the one in the proof of the classical convergence bound for the CG method, which predicts linear convergence in terms of the condition number of A: there the spectrum is also replaced by its convex hull. However, for bounding $\operatorname{dist}(\lambda_1, \Theta)$ it is only necessary that s_{m-1} is small on the discrete set $\{\lambda_2, \ldots, \lambda_N\}$. The optimal polynomials for both tasks can look quite different, see Figure 8.1 for a simple example. Therefore a precise upper bound for $\operatorname{dist}(\lambda_k, \Theta)$ needs to incorporate the fine structure of the spectrum, see [Bec00a, HKV05, Kui00].

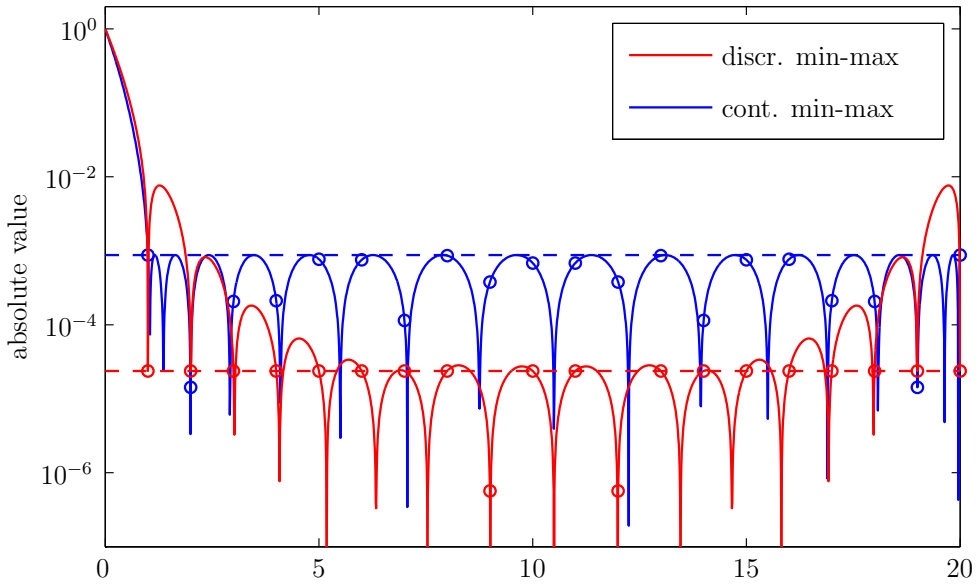


Figure 8.1: The absolute value of two polynomials of degree 17 taking the value 1 at 0. One polynomial (solid red) is minimal on the discrete set $\{1, 2, \ldots, 20\}$, and the other (solid blue) is minimal on the interval $[1, 20]$. The maximal absolute value attained by these polynomials on their respective sets of optimality is indicated by the dashed lines.

8.2 Asymptotic Distribution of Ritz Values

Let us first recall some recent asymptotic results about polynomial Ritz values. In the following we use notions from logarithmic potential theory, such as potential, energy and weak-star convergence, which were defined in Section 7.2.

The Polynomial Krylov Case. There is a rule of thumb proposed by Trefethen & Bau [TB97, p. 279] that $\operatorname{dist}(\lambda_k, \Theta)$ is small for eigenvalues λ_k in regions of "too little charge" for an equilibrium distribution. It was Kuijlaars [Kui00] who first quantified this rule in an asymptotic sense using logarithmic potential theory (see also [Bec00a, HKV05]). Since it is not possible to consider asymptotics for a single matrix A, we need a sequence of Hermitian matrices $A^{(N)} \in \mathbb{C}^{N\times N}$ with a joint eigenvalue distribution described by a probability measure σ. To be more precise, we denote by Λ_N the set of eigenvalues $\lambda_1^{(N)} < \cdots < \lambda_N^{(N)}$ of $A^{(N)}$ and let

$$\delta_N(\Lambda_N) := \frac{1}{N} \sum_{x\in\Lambda_N} \delta_x$$

be the normalized counting measure of the eigenvalues. We then assume a weak-star convergence $\delta_N(\Lambda_N) \xrightarrow{*} \sigma$. Sequences of matrices having a joint eigenvalue distribution occur quite frequently in applications, the most prominent examples being finite sections of Toeplitz operators, see for instance [BS99]. Matrices obtained by finite-difference or finite-element discretization of PDEs with varying mesh width can also have a joint eigenvalue distribution after suitable rescaling [BC07].

In addition to the above we require a sequence of vectors $\boldsymbol{b}^{(N)} \in \mathbb{C}^N$ and a fixed number $t \in (0, 1)$. Associated with $A^{(N)}$ and $\boldsymbol{b}^{(N)}$ are polynomial Ritz values of order $m = \lceil tN \rceil$, which we denote by $\theta_1^{(N)} < \cdots < \theta_m^{(N)}$ and collect in the set Θ_N. Under mild assumptions Kuijlaars showed that the distribution of these Ritz values is given by a measure μ_t in the sense that $\delta_N(\Theta_N) \xrightarrow{*} \mu_t$, where μ_t solves a *constrained equilibrium problem* from logarithmic potential theory. More precisely, μ_t is the unique minimizer of the (mutual) logarithmic energy

$$I(\mu) = I(\mu, \mu), \quad I(\mu_1, \mu_2) = \iint \log \frac{1}{|x - y|} \, \mathrm{d}\mu_1(x) \, \mathrm{d}\mu_2(y) \tag{8.3}$$

among all positive Borel measures μ of total mass $\|\mu\| = t$ satisfying $\mu \le \sigma$. The minimal energy property (8.3) is a consequence of the fact that Ritz values are zeros of discrete orthogonal polynomials and hence optimal in a certain sense (cf. Lemma 8.1). This relationship allows one to make use of weak asymptotics for discrete polynomials due to Rakhmanov [Rak96], Dragnev & Saff [DS97], Van Assche & Kuijlaars [KV99], Beckermann [Bec00b], and others [KR98, BR99, CV05]. The constraint $\mu \le \sigma$ results from

the fact that Ritz values are nowhere denser than eigenvalues, which follows from the well-known *interlacing property* (cf. [Par98, Theorem 10.1.1]):

$$\text{In every open interval } (\theta_j, \theta_{j+1}) \text{ there is at least one eigenvalue of } A. \tag{8.4}$$

We conclude that in parts of the real line where the constraint is active there are asymptotically as many Ritz values as eigenvalues. Let F_t be the maximum of the logarithmic potential

$$U^{\mu_t}(x) = \int \log \frac{1}{|x-y|} \, \mathrm{d}\mu_t(y)$$

on the real line (which is also the maximum in the complex plane) and define the set $\Sigma_t = \{x \in \mathbb{R} : U^{\mu_t}(x) = F_t\}$. Points outside Σ_t are of "too little charge" for an equilibrium distribution and Kuijlaars proved that eigenvalues lying in a neighborhood of a point $x \in \mathbb{R} \setminus \Sigma_t \subseteq \mathbb{R} \setminus \operatorname{supp}(\sigma - \mu_t)$ are approximated by Ritz values with a geometric rate. The close connection between charge distributions and logarithmic potentials allows us to think of Ritz values as electrons placed on A's spectral interval, moved to electrostatic equilibrium (the state of minimal energy (8.3)) and satisfying the interlacing property (8.4). Let us give an illustrative example.

Example 8.2. We consider a diagonal matrix $A \in \mathbb{R}^{N \times N}$ with equidistant eigenvalues in $[-1, 1]$ and a vector $\boldsymbol{b} \in \mathbb{R}^N$ with all entries being one, $N = 100$. The eigenvalue density is $\mathrm{d}\sigma / \mathrm{d}x = 1/2$. In Figure 8.2 (top) we show the density of the constrained equilibrium measure μ_t of total mass $t = 0.75$ and the associated potential U^{μ_t}, which attains its maximal value

$$F_t = t + t \cdot \log \frac{2}{\sqrt{1-t^2}} - \frac{1}{2} \log \frac{1+t}{1-t} \approx 0.607$$

on the interval $\Sigma_t = [-\sqrt{1-t^2}, \sqrt{1-t^2}]$ (these expressions are given in [Rak96]). We expect that the Ritz values of order $m \geq 75$ start converging to eigenvalues of A that are outside Σ_t. This is confirmed in the lower figure, where we plot the Ritz values of all orders $m = 1, \ldots, 100$ and the endpoints of Σ_t (black parabola) for varying $t = m/N$. We use different colors, listed in Table 8.1, to indicate the distance of each Ritz value to a closest eigenvalue of A. As expected, the Ritz values start converging to the extremal eigenvalues first since these have a low density compared to the (unconstrained) equilibrium measure of the interval $[-1, 1]$. Although the above statements are only of an asymptotic nature, they can obviously provide good predictions for matrices of moderate size.

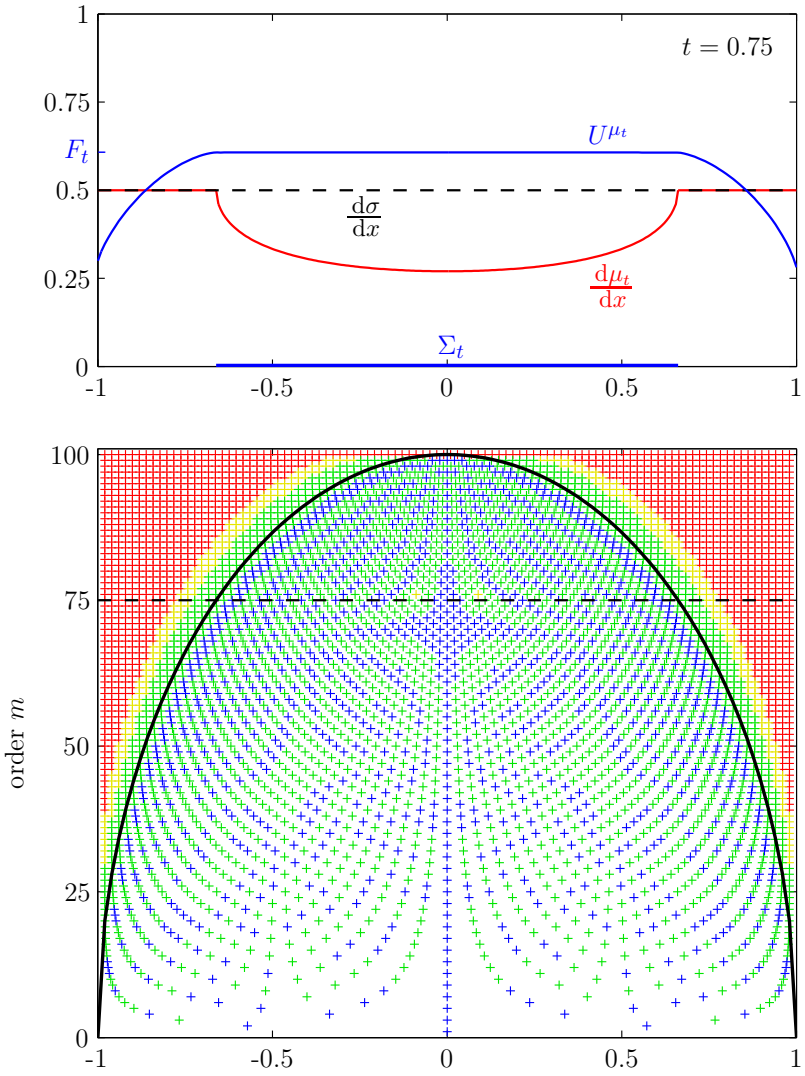


Figure 8.2: Constrained equilibrium problem and its connection to the convergence of polynomial Ritz values for a matrix with $N = 100$ equidistant eigenvalues in $[-1, 1]$. The black parabola in the lower figure indicates the endpoints of the interval Σ_t ($t = m/N$), outside of which the Ritz values have converged.

Color	Distance of a Ritz value θ to the spectrum
Red	$\mathrm{dist}(\theta, \Lambda(A_N)) < 10^{-7.5}$
Yellow	$10^{-7.5} \leq \mathrm{dist}(\theta, \Lambda(A_N)) < 10^{-5.0}$
Green	$10^{-5.0} \leq \mathrm{dist}(\theta, \Lambda(A_N)) < 10^{-2.5}$
Blue	$10^{-2.5} \leq \mathrm{dist}(\theta, \Lambda(A_N))$

Table 8.1: Color codes for Ritz values

The Rational Krylov Case. Generalizing the works of Kuijlaars [Kui00] and Beckermann [Bec00a], asymptotic results for rational Ritz values were obtained in [BGV10]. In this case one needs to take into account the influence of the poles $\xi_1^{(N)}, \dots, \xi_{m-1}^{(N)}$, which are collected in a multiset Ξ_N and assumed to have a weak-star limit $\delta_N(\Xi_N) \xrightarrow{*} \nu_t$ (this measure counts poles according to their multiplicity). These poles cause technical difficulties when lying in the spectrum of A, for example, and the major part of [BGV10] deals with this case. This critical situation is primarily of interest for eigenvalue computations, where a shift is often placed in proximity of a sought-after eigenvalue. In the context of approximating matrix functions $f(A)\boldsymbol{b}$, however, the poles are typically away from the numerical range $\mathbb{W}(A)$. To keep the exposition simple we therefore limit ourselves to the case of strict separation between eigenvalues and poles.

Assumption 1: There exist disjoint compact sets Λ and Ξ such that for all N there holds $\Lambda_N \subset \Lambda$ and $\Xi_N \subset \Xi$.

Two more technical assumptions are necessary. Logarithmic potential theory provides asymptotics in an Nth root sense and hence the eigenvalues Λ_N should not cluster exponentially (this situation could not be resolved by these asymptotics). The following assumption prevents exponential clustering, but still allows for equidistant eigenvalues, Chebyshev eigenvalues (the eigenvalues of the 1D Laplacian), and more general sets of points [DS97]. It also guarantees that U^σ is continuous [BGV10, Lemma A.4].

Assumption 2: For any sequence $\Lambda_N \ni \lambda^{(N)} \to \lambda$ for $N \to \infty$ there holds

$$\limsup_{\delta \to 0+} \limsup_{N \to \infty} \frac{1}{N} \sum_{0 < |\lambda_j^{(N)} - \lambda^{(N)}| \le \delta} \log \frac{1}{|\lambda_j^{(N)} - \lambda^{(N)}|} = 0.$$

Rational Ritz values Θ_N can only approximate eigenvalues whose associated eigenvectors are present in the starting vector $\boldsymbol{b}^{(N)}$. We therefore need to ensure that $\boldsymbol{b}^{(N)}$ has sufficiently large eigencomponents in all eigenvectors of $A^{(N)}$.

Assumption 3: The eigencomponents $w(\lambda_j^{(N)}) \in [0, 1]$ defined in (8.1) satisfy

$$\liminf_{N \to \infty} \min_j w(\lambda_j^{(N)})^{1/N} = 1.$$

We recall from above that polynomial Ritz values distribute like electrons placed on the spectral interval of A moving to electrostatic equilibrium and being nowhere denser than

the eigenvalues. In the rational Krylov case these electrons are attracted by positive charges (the poles of the rational Krylov space) whose distribution is described by the measure ν_t. The rational Ritz values also satisfy the interlacing property (8.4) because they are polynomial Ritz values for a modified starting vector. From this physical intuition it seems natural that the rational Ritz values distribute according to a measure μ_t minimizing the energy

$$I(\mu - \nu_t) = I(\mu) - 2I(\mu, \nu_t) + I(\nu_t) \geq 0$$

among all positive Borel measures μ of total mass $\|\mu\| = t$ satisfying the constraint $\mu \leq \sigma$. Since $I(\nu_t)$ may be infinite (e.g., if ν_t has mass points), it is more adequate to minimize $I(\mu) - 2I(\mu, \nu_t)$ instead. As in the polynomial case we expect that rational Ritz values start converging to eigenvalues in regions of the real line where the constraint is active. Indeed this happens with a geometric rate as asserted in the following theorem [BGV10, Thm. 3.1].

Theorem 8.3. *Under the above assumptions the rational Ritz values of order $m = \lceil tN \rceil$ have an asymptotic distribution described by $\delta_N(\Theta_N) \to \mu_t$, where μ_t is the unique minimizer of $I(\mu) - 2I(\mu, \nu_t)$ among all positive Borel measures μ of total mass $\|\mu\| = t$ satisfying $\mu \leq \sigma$.*

Define F_t as the maximum of $U^{\mu_t - \nu_t}$ in the whole complex plane and let $\Sigma_t = \{z \in \mathbb{C} : U^{\mu_t - \nu_t}(z) = F_t\}$. Let $J \subset \mathbb{R} \setminus \Sigma_t \subset \mathbb{R} \setminus \operatorname{supp}(\sigma - \mu_t)$ be a closed interval. Then all eigenvalue sequences $\{\lambda^{(N)} \in \Lambda_N\} \subset J$ with $\lambda^{(N)} \to \lambda$ for $N \to \infty$ satisfy

$$\lim_{N \to \infty} \operatorname{dist}(\lambda^{(N)}, \Theta_N)^{1/N} = \exp(2(U^{\mu_t - \nu_t}(\lambda) - F_t)),$$

with the possible exclusion of at most one unique "exceptional eigenvalue" in each set Λ_N.

Example 8.4. As in Example 8.2 we consider a diagonal matrix $A \in \mathbb{R}^{N \times N}$ with equidistant eigenvalues in $[-1, 1]$ and the vector $\boldsymbol{b} \in \mathbb{R}^N$ with all entries being one, $N = 100$. All poles of the rational Krylov space are at the point $\xi = 1.1$, so that in iteration m we have $\nu_t = t \cdot \delta_{1.1}$ with $t = m/N$. In Figure 8.2 (top) we show the density of the constrained equilibrium measure μ_t of total mass $t = 0.75$ and the associated potential $U^{\mu_t - \nu_t}$. We have computed μ_t by minimizing the energy of a measure with piecewise linear density using an active set method implemented in MATLAB's `quadprog`. We expect that the Ritz values of order $m \geq 75$ start converging to eigenvalues of A that are outside the set Σ_t. This is

confirmed in the lower figure, where we plot the Ritz values of all orders $m = 1, \ldots, 100$ and the endpoints of Σ_t (black curve) for varying $t = m/N$. We use the color codes of Table 8.1 to display the distance of each Ritz value to a closest eigenvalue of A. We observe that the rational Ritz values start converging to the right-most eigenvalues of A first, which are the ones closest to the pole ξ.

We note that by Theorem 5.13 the interpolation points of the shift-and-invert method with $A - \xi I$ were distributed according to the same measure μ_t.

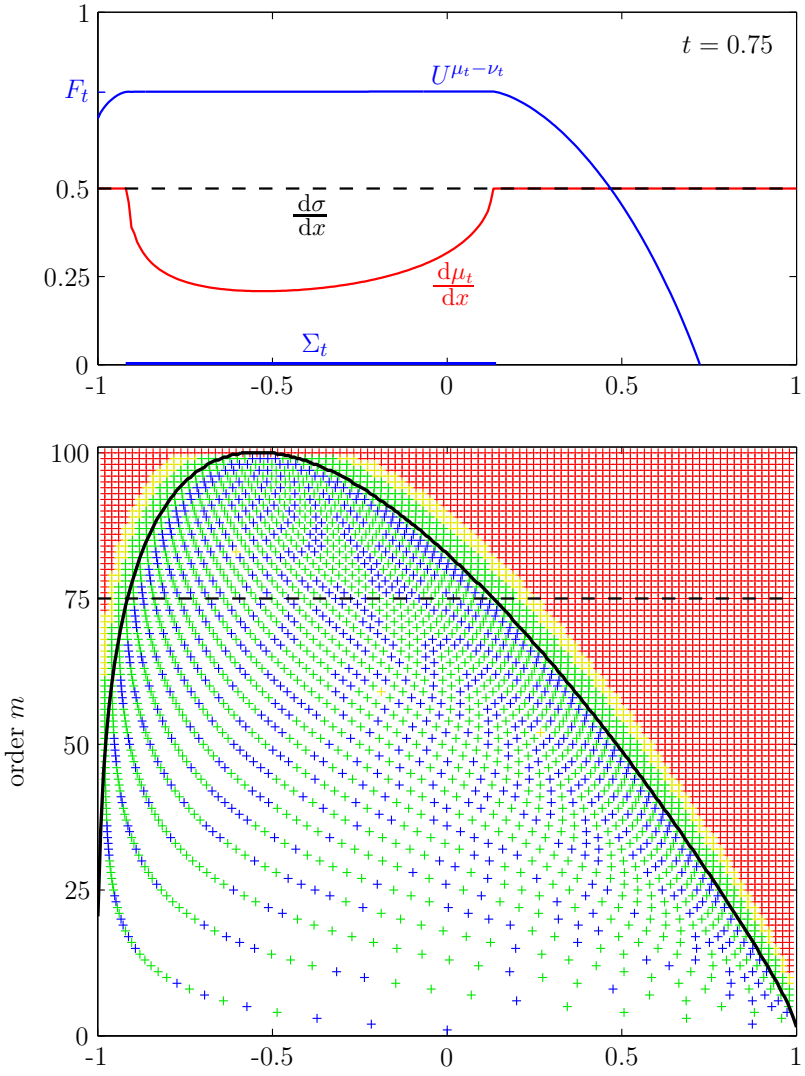


Figure 8.3: Constrained weighted equilibrium problem and its connection to the convergence of rational Ritz values for a matrix with 100 equidistant eigenvalues in $[-1, 1]$. All poles of the rational Krylov space are at the point $\xi = 1.1$. The black curve in the lower figure indicates the endpoints of the interval Σ_t ($t = m/N$), outside of which the rational Ritz values have converged.


8ratritz

8.3 Superlinear Convergence

The analytic description of the convergence of polynomial Ritz values is an important step towards understanding the superlinear convergence behavior of the CG method (see [BK01a, BK01b, BK02] and the reviews [DTT98, Kui06, Bec06]). Superlinear convergence can also be observed with discrete rational approximation and has been studied, e.g., in connection with the ADI method [BG10]. Also rational Rayleigh–Ritz approximations for matrix functions show superlinear convergence effects and in fact we have already seen them, for example in Figure 7.7 (top) on page 110, where we approximated the resolvent function $f^\tau(A)\boldsymbol{b}$, $f^\tau(z) = (z-\tau)^{-1}$, for a symmetric positive definite matrix A from a rational Krylov space with poles on the negative real axis. Note that the error curves belonging to small parameters close to $\tau = 1i$ start converging superlinearly after about 10 iterations and ultimately "overtake" the curves belonging to large parameters close to $\tau = 1000i$. In the following we study these effects with the help of a simple example.

We consider the symmetric Toeplitz matrix

$$A = \begin{bmatrix} q^0 & q^1 & q^2 & \\ q^1 & q^0 & q^1 & \ddots \\ q^2 & q^1 & q^0 & \ddots \\ & \ddots & \ddots & \ddots \end{bmatrix} \in \mathbb{R}^{N\times N}, \qquad \text{with } q \in (0,1).$$

It is known [KMS53, BK01b] that the asymptotic eigenvalue distribution for $N \to \infty$ is described by a measure σ supported on the positive interval $[\alpha, \beta]$ with density

$$\frac{\mathrm{d}\sigma}{\mathrm{d}x}(x) = \frac{1}{\pi x\sqrt{(x-\alpha)(\beta-x)}}, \qquad \alpha = \frac{1-q}{1+q}, \qquad \beta = \frac{1+q}{1-q}.$$

One can show [BGV10, §4] that if all poles of the rational Krylov space are at a point $\xi > \beta$, i.e., $\nu_t = t \cdot \delta_\xi$, then the set Σ_t from Theorem 8.3 is the interval $\Sigma_t = [\alpha, b(t)]$ with

$$b(t) = \min\left\{\beta, \frac{\xi}{t^2\beta(\xi-\alpha)+1}\right\}.$$

We choose $N = 100$ and $q = 1/3$ so that the matrix A has eigenvalues in $\Sigma = [1/2, 2]$. In Figure 8.4 (top) we illustrate the convergence of the rational Ritz values obtained with a rational Krylov space with all poles in $\xi = 2.5$ and a random starting vector $\boldsymbol{b}$

with normally distributed entries. The black curve is the right endpoint of Σ_t plotted for varying $t \in (0,1)$. This curve predicts that the rational Ritz values of order $m \geq 25$ start converging to the right-most eigenvalues of A, which is in good agreement with the observed convergence.

Let us approximate the resolvent function $f^\tau(A)\boldsymbol{b}$, $f^\tau(z) = (z-\tau)^{-1}$, from the same rational Krylov space for the three parameters

$$\tau_1 = 0, \quad \tau_2 = 0.32, \quad \tau_3 = 2.01,$$

which are chosen such that with the technique described in Section 7.4 we predict linear convergence at rates

$$R_1 = 1.67, \quad R_2 = 1.39, \quad R_3 = 1.39,$$

respectively. The solid lines in Figure 8.4 (bottom) show the corresponding convergence curves, which look approximately linear for orders $m < 25$. For $m \geq 25$, that is, when the right-most rational Ritz values start converging to eigenvalues of A, we observe superlinear convergence. An intuitive explanation for this behavior goes as follows: by Theorem 4.8 we know that the rational function r_m^τ underlying the Rayleigh–Ritz approximation $\boldsymbol{f}_m^\tau = r_m^\tau(A)\boldsymbol{b}$ interpolates the function f^τ at the rational Ritz values $\theta_1, \ldots, \theta_m$. Since these Ritz values approximate very well the eigenvalues outside the interval $\Sigma_t = [1/2, b(t)]$, $t = m/N$, and hence the interpolation error at these eigenvalues is already small, the near-best Rayleigh–Ritz extraction puts most of its "effort" into reducing the error on the remaining set Σ_t. The near-best approximation problem thus takes place on shrinking sets Σ_t. If $R(t,\tau)$ denotes the convergence rate of rational best approximation for f^τ on Σ_t with all poles in ξ, then we expect (similar to the reasoning in Section 7.5.3) that the Rayleigh–Ritz approximation of order m reduces the initial error at least by the factor

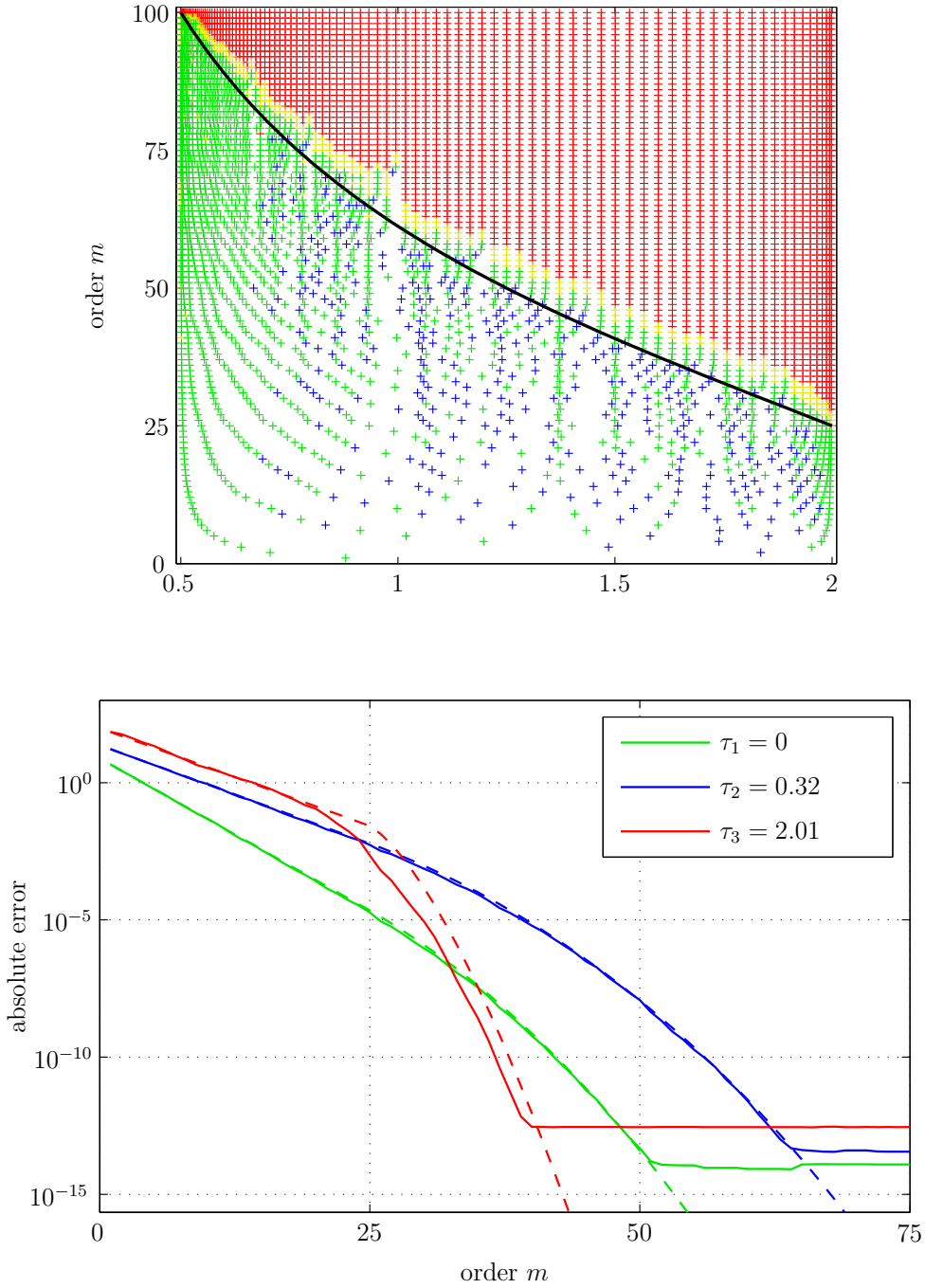
$$E_m(\tau) = \left[R(1/N,\tau)R(2/N,\tau)\cdots R(m/N,\tau)\right]^{-1}, \tag{8.5}$$

and this quantity is shown in Figure 8.4 (bottom) for the parameters τ_1, τ_2, τ_3 as dashed lines. The convergence acceleration clearly depends on the location of the parameter τ relative to the set Σ_t. For example, the parameters τ_1 and τ_2 benefit less from superlinear convergence than τ_3, because they lie to the left of Σ_t and the rational Ritz values converge on the other side. Initially, that is, for $m < 25$, the error curves for τ_2 and τ_3 are

parallel because we have chosen these parameters such that the corresponding asymptotic convergence rates are equal. However, for $m \geq 25$ this linear rate would be a quite pessimistic prediction for the actual convergence. All these effects are reflected in the quantity (8.5). Note that

$$[E_m(\tau)]^{1/N} = \exp\left(-\frac{1}{N}\sum_{j=1}^{m}\log R(j/N,\tau)\right) \to \exp\left(-\int_0^t \log R(s,\tau)\,\mathrm{d}s\right)$$

as $N \to \infty$ and $m/N \to t$. In the polynomial Krylov case we have $R(t,\tau) = \exp(g_{\mathbb{C}\setminus\Sigma_t}(\tau))$, where $g_{\mathbb{C}\setminus\Sigma_t}$ denotes the Green's function of $\mathbb{C}\setminus\Sigma_t$, and if $\tau = 0$ we recover the asymptotic error formula for the CG method derived more rigorously in [BK01b, Thm. 2.1].



8superlin

Figure 8.4: The convergence of rational Ritz values (top) yields superlinear convergence of Rayleigh–Ritz approximations for the resolvent function $f^\tau(A)\boldsymbol{b}$, $f^\tau(z) = (z - \tau)^{-1}$. The convergence acceleration depends on the location of the parameter τ relative to the set Σ_t of unconverged Ritz values.

9 Numerical Experiments

Why does MATLAB
have a `sin` *function,*
but no `forgive` *function?*
P. J. Acklam

In this chapter we test and illustrate various aspects of rational Krylov methods by numerical examples. These examples include Maxwell's equations in the time and frequency domain, the approximation of the sign function of a QCD matrix, an advection–diffusion problem, and the solution of a wave equation. All computations are done in MATHWORKS MATLAB (release 2008b) on a quad-core AMD OPTERON processor running SUSE LINUX ENTERPRISE SERVER (version 10). The FEM discretizations are generated by the COMSOL MULTIPHYSICS package (version 3.5). In the first two examples we use the direct solver PARDISO (version 3.2), which is invoked through a MEX-interface from MATLAB and allows one to reuse the information of the analysis step if systems with identical nonzero structure are solved.

9.1 Maxwell's Equations

9.1.1 Time Domain

The modeling of transient electromagnetic fields in inhomogeneous media is a typical task arising, for example, in geophysical prospecting [OH84, GHNS86, DK94]. Such models

can be based on the quasi-static Maxwell's equations

$$\begin{aligned} \nabla \times \boldsymbol{e} + \mu \partial_\tau \boldsymbol{h} &= \boldsymbol{0}, \\ \nabla \times \boldsymbol{h} - \sigma \boldsymbol{e} &= \boldsymbol{j}^e, \\ \nabla \cdot \boldsymbol{h} &= 0, \end{aligned} \tag{9.1}$$

where

$\boldsymbol{e} = \boldsymbol{e}(\boldsymbol{x}, \tau)$	is the electric field,
$\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{x}, \tau)$	is the magnetic field,
$\sigma = \sigma(\boldsymbol{x})$	is the electric conductivity,
$\mu = 4\pi \cdot 10^{-7}$	is the magnetic permeability, and
$\boldsymbol{j}^e = \boldsymbol{j}^e(\boldsymbol{x}, \tau)$	is the external source current density.

The variables $\boldsymbol{x} = [x, y, z]^T$ and τ correspond to space and time, respectively. In geophysics the plane $z = 0$ is the earth–air interface and z increases downwards. After eliminating $\boldsymbol{h}$ from (9.1) we obtain the second order partial differential equation

$$\nabla \times \nabla \times \boldsymbol{e} + \mu\sigma \partial_\tau \boldsymbol{e} = -\mu \partial_\tau \boldsymbol{j}^e \tag{9.2}$$

for the electric field. The source term $\boldsymbol{j}^e$ typically results from a known stationary transmitter with a driving current that is shut off at time $\tau = 0$, i.e.,

$$\boldsymbol{j}^e(\boldsymbol{x}, \tau) = \boldsymbol{q}(\boldsymbol{x}) H(-\tau) \tag{9.3}$$

with the vector field $\boldsymbol{q}$ denoting the spatial current pattern and the Heaviside unit step function H.

To reduce the problem to two space dimensions we assume now that σ and $\boldsymbol{j}^e$ are invariant in the y-direction and the vector field $\boldsymbol{j}^e(\boldsymbol{x}, \tau) = [0, j^e(x, z, \tau), 0]^T$ points only into the y-direction. In this case the electric field $\boldsymbol{e}(\boldsymbol{x}, \tau) = [0, e(x, z, \tau), 0]^T$ has only one component and hence is divergence free. Using the identity $\nabla \times \nabla \times \boldsymbol{e} = \nabla(\nabla \cdot \boldsymbol{e}) - \nabla^2 \boldsymbol{e}$ in (9.2), we

arrive at a scalar bidimensional heat equation for $e = e(x, z, \tau)$,

$$-\nabla^2 e + \mu\sigma\partial_\tau e = -\mu\partial_\tau j^e. \tag{9.4}$$

To restrict this equation to the region of interest $z > 0$ (which is the earth), we impose an exact boundary condition at the earth–air interface of the form

$$\partial_z e = Te + \mu\partial_\tau j^e, \quad \text{for } z = +0,$$

where T is a linear nonlocal convolution operator in x given in [GHNS86]. This condition ensures that the tangential component of e is continuous across the line $z = 0$ and that the Laplace equation $-\nabla^2 e = 0$ is satisfied for $z < 0$. As spatial domain we consider a rectangle $\Omega = (-10^5, 10^5) \times (0, 4000)$ and we assume, in addition to the exact boundary condition on the top, homogeneous Dirichlet data on the lower, left and right boundary of Ω. The structure of the conductivity σ in our model is sketched in Figure 9.1 (top). We assume that for times $\tau < 0$ the source term j^e corresponds to a steady current in a double line source located on the earth–air interface, and that this source is shut off at time $\tau = 0$. The thereby induced electric field at time $\tau_0 = 10^{-6}$ is assigned as initial value e_0 for (9.4) (in the short time interval $[0, \tau_0]$ the electric field does not reach any inhomogeneous conductivity structures of our model and is known analytically [OH84], cf. Figure 9.1 (middle)). Under these assumptions the discretization of (9.4) using linear finite elements on triangles yields a linear ordinary differential equation

$$M\boldsymbol{e}'(\tau) = K\boldsymbol{e}(\tau), \quad \boldsymbol{e}(\tau_0) = \boldsymbol{e}_0$$

with symmetric matrices K, $M \in \mathbb{R}^{N\times N}$ and vectors $\boldsymbol{e}_0$, $\boldsymbol{e}(\tau) \in \mathbb{R}^N$ ($N = 20134$). The solution of this problem is explicitly given as

$$\boldsymbol{e}(\tau) = f^\tau(A)\boldsymbol{b}, \quad \text{where } f^\tau(z) = e^{(\tau-\tau_0)z},\ A = M^{-1}K,\ \boldsymbol{b} = \boldsymbol{e}_0.$$

Note that the matrix A is in general not symmetric, but its eigenvalues are real since it is similar to $M^{-1/2}KM^{-1/2}$. In fact, we could symmetrize the problem using the identity

$$f^\tau(M^{-1}K) = M^{-1/2} f^\tau(M^{-1/2}KM^{-1/2})M^{1/2},$$

but this requires additional operations with the matrices $M^{-1/2}$ and $M^{1/2}$. Therefore we prefer not to do this symmetrization and no complications will arise when (formally) working with A. Since

$$(I - A/\xi)^{-1}A = (I - M^{-1}K/\xi)^{-1}M^{-1}K = (M - K/\xi)^{-1}K,$$

we have to solve linear systems with the matrix $M - K/\xi$, which is still sparse since K and M have similar nonzero patterns and also symmetric if the pole ξ is real.

A computational task, arising for example with inverse problems, is to model the transient behavior of the electric field by approximating $\boldsymbol{f}_m^\tau$ for many time parameters $\tau \in T$. In our example T contains 25 logspaced points in the interval $[\tau_0 = 10^{-6}, 10^{-3}]$. An approximation $\boldsymbol{f}_m^\tau$ is considered accurate enough if $\|f^\tau(A)\boldsymbol{b} - \boldsymbol{f}_m^\tau\| \leq 10^{-8}$. The exact solution $f^\tau(A)\boldsymbol{b}$ is computed by evaluating the best uniform rational approximation of type $(16, 16)$ to the exponential function on $(-\infty, 0]$ obtained by the Carathéodory–Fejér method (see [TWS06, Fig. 4.1]). A Carathéodory–Fejér approximation of type $(9, 9)$ would actually suffice to achieve the desired stopping accuracy of 10^{-8}. This rational function has 4 complex conjugate poles so that we would need to solve $(4+1) \cdot 25 = 125$ (complex) linear systems of equations, which takes PARDISO about 15 seconds if the analysis step is done exactly once for all systems.

We test two different sequences of *real* poles for the rational Krylov space $\mathcal{Q}_m$. The first sequence contains 3 cyclically repeated poles computed for the spectral inclusion set $\Sigma = [-\infty, 0]$ by the technique described in Section 7.6.1 (see Example 7.14 on page 114). Corollary 7.13 tells us that at most $m = 70$ iterations are required to achieve an absolute accuracy of 10^{-8}. The second sequence are the "Zolotarev poles" given in [DKZ09] (cf. Section 7.6.2). This sequence requires information about the spectral interval of A, which we estimated by a few iterations of the rational Arnoldi algorithm as $\Lambda(A) \subseteq [-10^8, -1]$. By (7.20) we expect an asymptotic convergence rate $R = 1.28$.

Figure 9.2 shows the convergence curves of Rayleigh–Ritz approximations extracted from rational Krylov spaces with the two different pole sequences. The number of subroutine calls making up the essential part of the computation and the overall time spent in each subroutine are given in Table 9.1. The timings are based on the reports of MATLAB's profiler tool and averaged over ten runs of the rational Arnoldi algorithm. In both cases

we use identical implementations of this algorithm, the only difference being the pole sequences. The analysis step of Pardiso is done exactly once because the nonzero structure of $M - K/\xi$ is independent of ξ. The advantage of cyclically repeated poles becomes obvious in the time spent with factorizing these matrices. After obtaining the reduced rational Arnoldi decomposition $AV_mK_m = V_{m+1}\underline{H_m}$ with $\|\boldsymbol{b}\|V_m\boldsymbol{e}_1 = \boldsymbol{b}$, the sought-after Rayleigh–Ritz approximation is computed as $\boldsymbol{f}_m^\tau = V_m f^\tau(H_m K_m^{-1})\|\boldsymbol{b}\|\boldsymbol{e}_1$. Here we do not make any use of orthogonality and hence no reorthogonalization is required in the rational Arnoldi algorithm and the number of inner products and vector-vector sums (in Table 9.1 subsumed under "orthogonalization steps") is $m(m+1)/2$. Note that the spectral interval of A is already large enough so that the cyclically repeated poles also outperform the Zolotarev poles in terms of required iterations m.

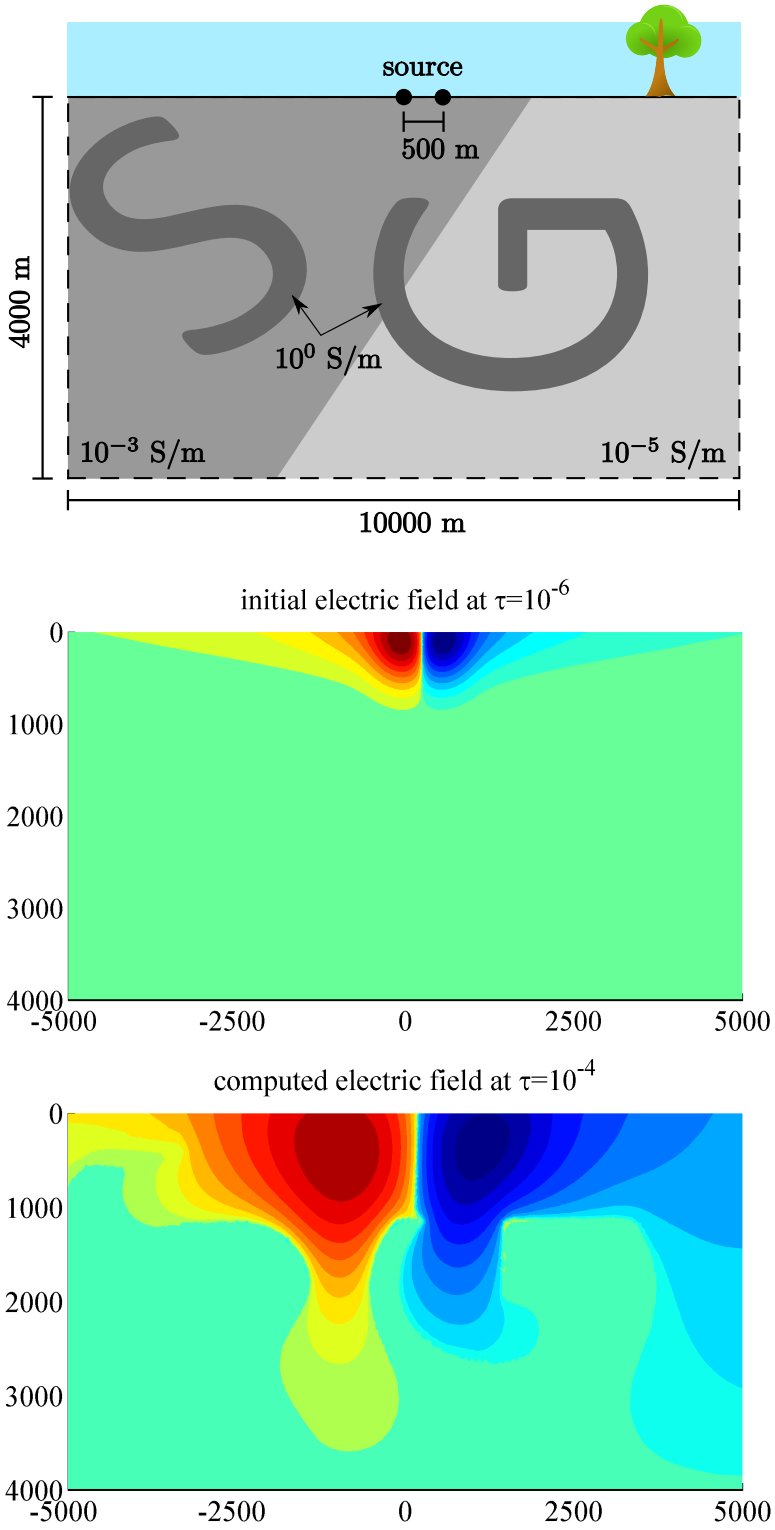


Figure 9.1: On the top is sketched a cutout of the spatial domain showing the conductivity structure and the location of the double line source on the earth–air interface. The other two pictures show snapshots of the electric field at initial time $\tau_0 = 10^{-6}$ and at $\tau = 10^{-4}$.

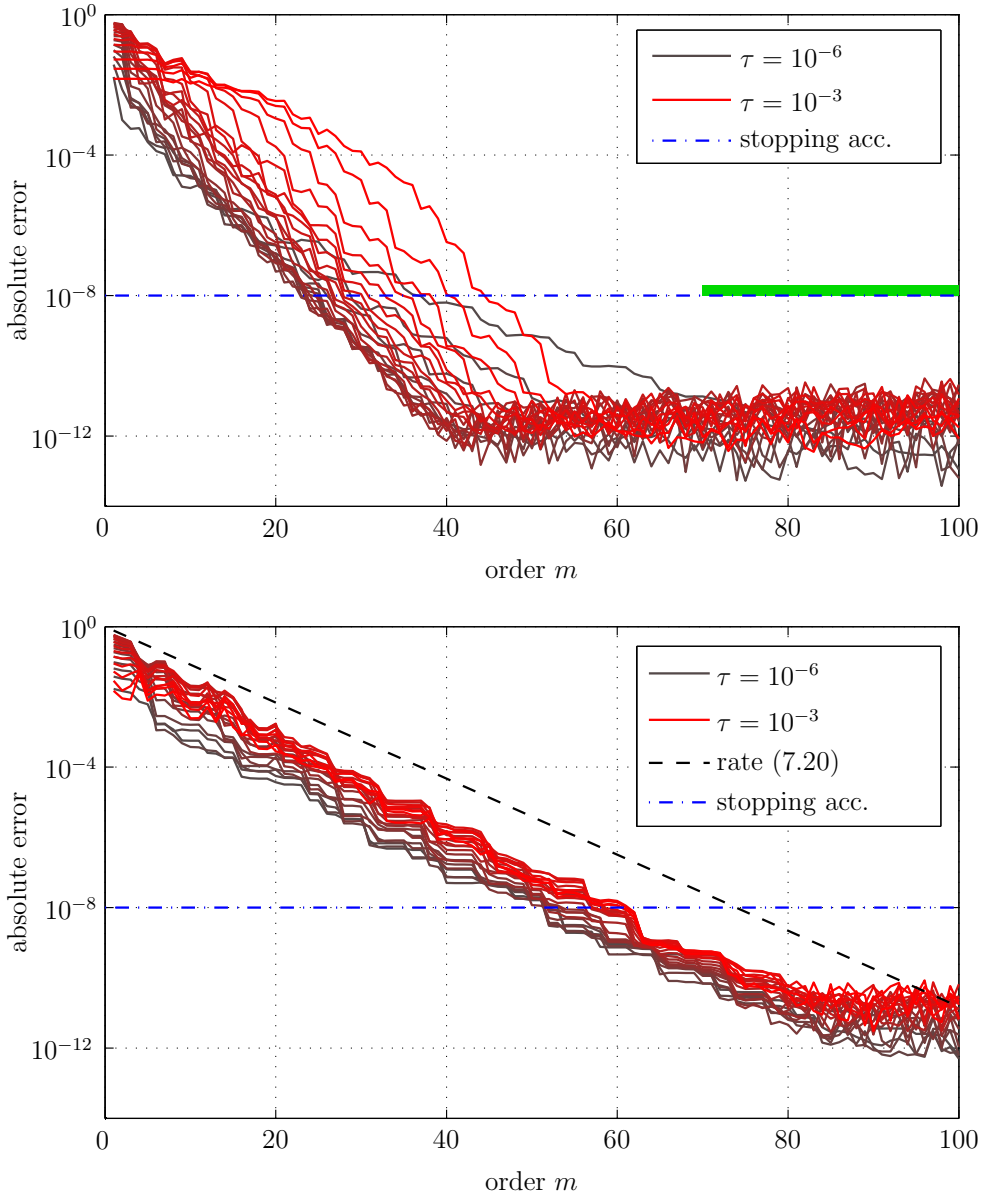


Figure 9.2: Error curves of Rayleigh–Ritz approximations for $f^\tau(A)\boldsymbol{b} = e^{(\tau-\tau_0)A}\boldsymbol{b}$ extracted from a rational Krylov space with cyclically repeated poles (top) and Zolotarev poles (below).

	Cyclically repeated poles	Zolotarev poles
Iterations m	45	62
PARDISO analysis	1 (148 ms)	1 (148 ms)
PARDISO factorizations	3 (238 ms)	62 (4920 ms)
PARDISO solves	45 (325 ms)	62 (448 ms)
Orthogonalization steps	1035 (282 ms)	1953 (533 ms)
Compute $\boldsymbol{f}_m^\tau$	25 (203 ms)	25 (288 ms)
Total	(1196 ms)	(6337 ms)

Table 9.1: Number of subroutine calls and timings. The problem size is $N = 20134$.

9.1.2 Frequency Domain

Applying the Fourier transform

$$\hat{\boldsymbol{e}}(\boldsymbol{x},\omega) = \int_{-\infty}^{\infty} \boldsymbol{e}(\boldsymbol{x},t)\exp(-i\omega t)\,\mathrm{d}t, \quad \omega \in \mathbb{R},$$

to (9.2) and (9.3), using the fact that $H(-\tau)$ transforms to multiplication by $-1/(i\omega)$, we obtain the equation

$$\nabla \times \nabla \times \hat{\boldsymbol{e}} + i\omega\mu\sigma\hat{\boldsymbol{e}} = \mu\boldsymbol{q}. \tag{9.5}$$

A common approach used in geophysics is to compute $\hat{\boldsymbol{e}}(\boldsymbol{x},\omega)$ for many frequencies ω in a real interval $[\omega_{\min},\omega_{\max}]$ and to synthesize the time-domain solution by inverse Fourier transform [NHA86]. As spatial domain we consider a cube with side-lengths 1000 having a layered conductivity structure with the top layer corresponding to air (cf. Figure 9.3). The finite-element discretization of (9.5) reads as

$$K\hat{\boldsymbol{e}} + i\omega M\hat{\boldsymbol{e}} = \mu\boldsymbol{q}$$

with symmetric matrices K, $M \in \mathbb{R}^{N\times N}$ and vectors $\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}(\omega)$, $\boldsymbol{q} \in \mathbb{R}^N$ ($N = 67937$), and the solution is given in terms of the resolvent function

$$\hat{\boldsymbol{e}}(\omega) = f^{\tau}(A)\boldsymbol{b}, \quad \text{where } f^{\tau}(z) = (z-\tau)^{-1},\ \tau = -i\omega,\ A = M^{-1}K,\ \boldsymbol{b} = \mu M^{-1}\boldsymbol{q}.$$

The parameters τ are elements of an *imaginary* parameter interval $T = i[\omega_{\min},\omega_{\max}]$. In our example we choose $T = i[10^3, 10^9]$ and approximate $f^{\tau}(A)\boldsymbol{b}$ for 25 logspaced parameters $\tau \in T$. Again we test two different sequences of poles. The first sequence consists of *real* poles computed by the method introduced in Section 7.5.3, i.e., these poles are obtained by successively maximizing the asymptotic convergence rate $\widetilde{R}(\tau)$ expected for all parameters $\tau \in T$. The resulting overall convergence rate is the minimum of the geometric mean of all single convergence rates. The second sequence are *imaginary* Zolotarev poles in T (cf. Section 7.5.2). For the computation of both sequences we assume that $\Lambda(M^{-1}K) \subset \Sigma := (-\infty, 0]$.

It is surprising that the real poles cluster in a few points, e.g., among 100 computed poles the values 10^3 and 10^9 (which are asymptotically optimal for the extremal parameters in T)

are repeated for 22 times each, cf. Figure 9.4 (top). To reduce the number of factorizations it seems reasonable to group the poles occuring in clusters. To this end we partition the interval $[\xi_{\min}, \xi_{\max}]$ containing the real poles ξ_j in 100 subintervals of geometrically increasing length and replace the ξ_j lying in one subinterval by their geometric mean ($j = 1, \ldots, m$). This way we can reduce the number of distinct poles from 100 to 13, without noticeable degradation of convergence. Such a reduction seems not possible for the imaginary Zolotarev poles since these are spread all over the parameter interval T, see Figure 9.4 (bottom).

In Figure 9.5 we show the residual curves of Rayleigh–Ritz approximations computed with identical implementations of the rational Arnoldi algorithm. As expected, the method with Zolotarev poles on the imaginary axis converges at a faster rate $R = 1.45$ than the method with real poles, which converges at rate $R = 1.27$. However, considering the computation times given in Table 9.2 it becomes obvious that complex arithmetic should be avoided: compared to using the real pole sequence the method with imaginary Zolotarev poles requires more than the 8-fold total computation time. The accuracy loss of about 5 digits indicates that the problem we are solving is very sensitive to perturbations. This phenomenon should be subject of future research. First experiments indicate that appropriate scalings of K and M, known to reduce the sensitivity of generalized eigenproblems [Bet08], improve the final accuracy by 1 or 2 digits.

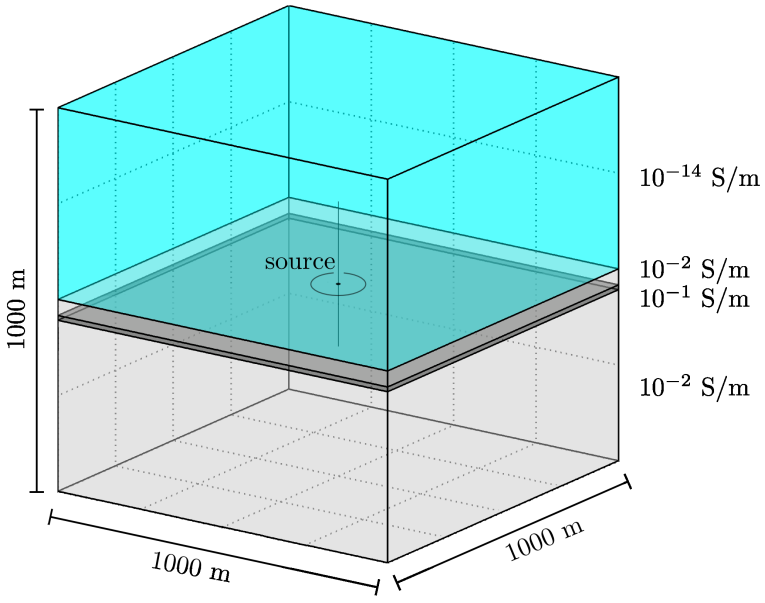


Figure 9.3: Computational domain and conductivity structure for the 3D Maxwell problem in the frequency domain.

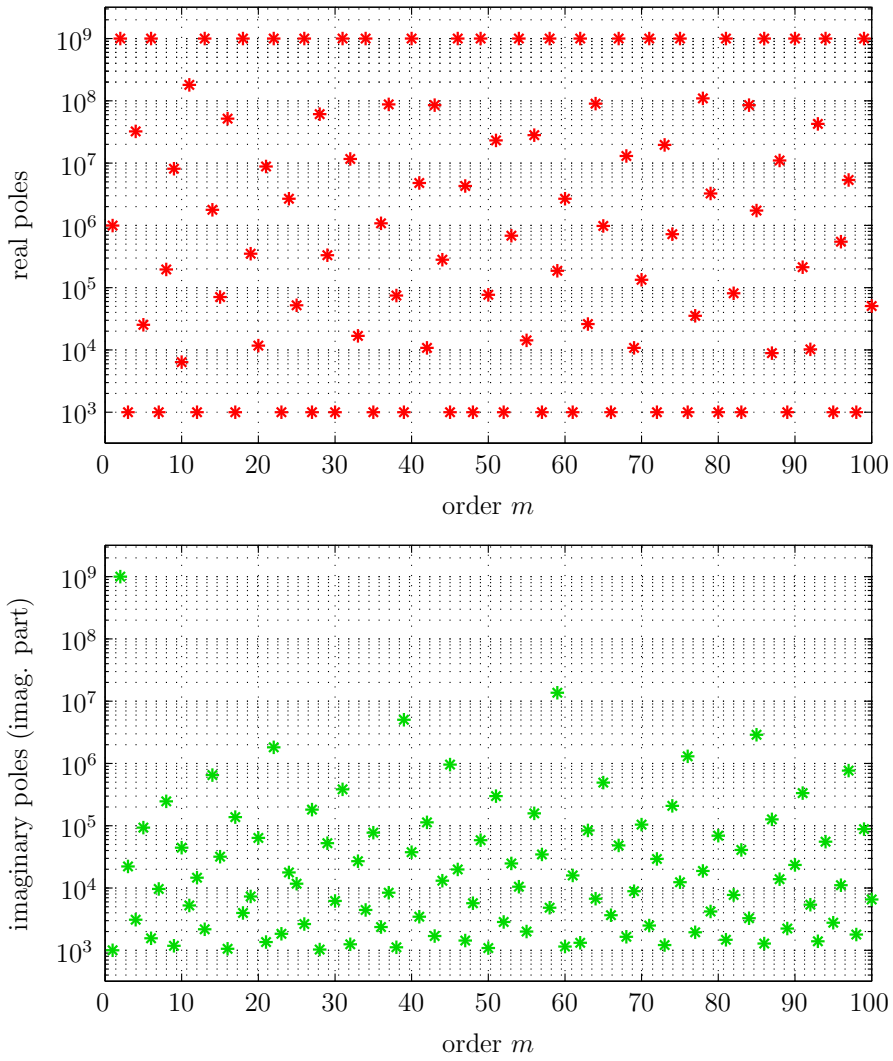


Figure 9.4: Sequences of real poles obtained by successive minimization of the asymptotic convergence rate (top) and imaginary Zolotarev poles (below).

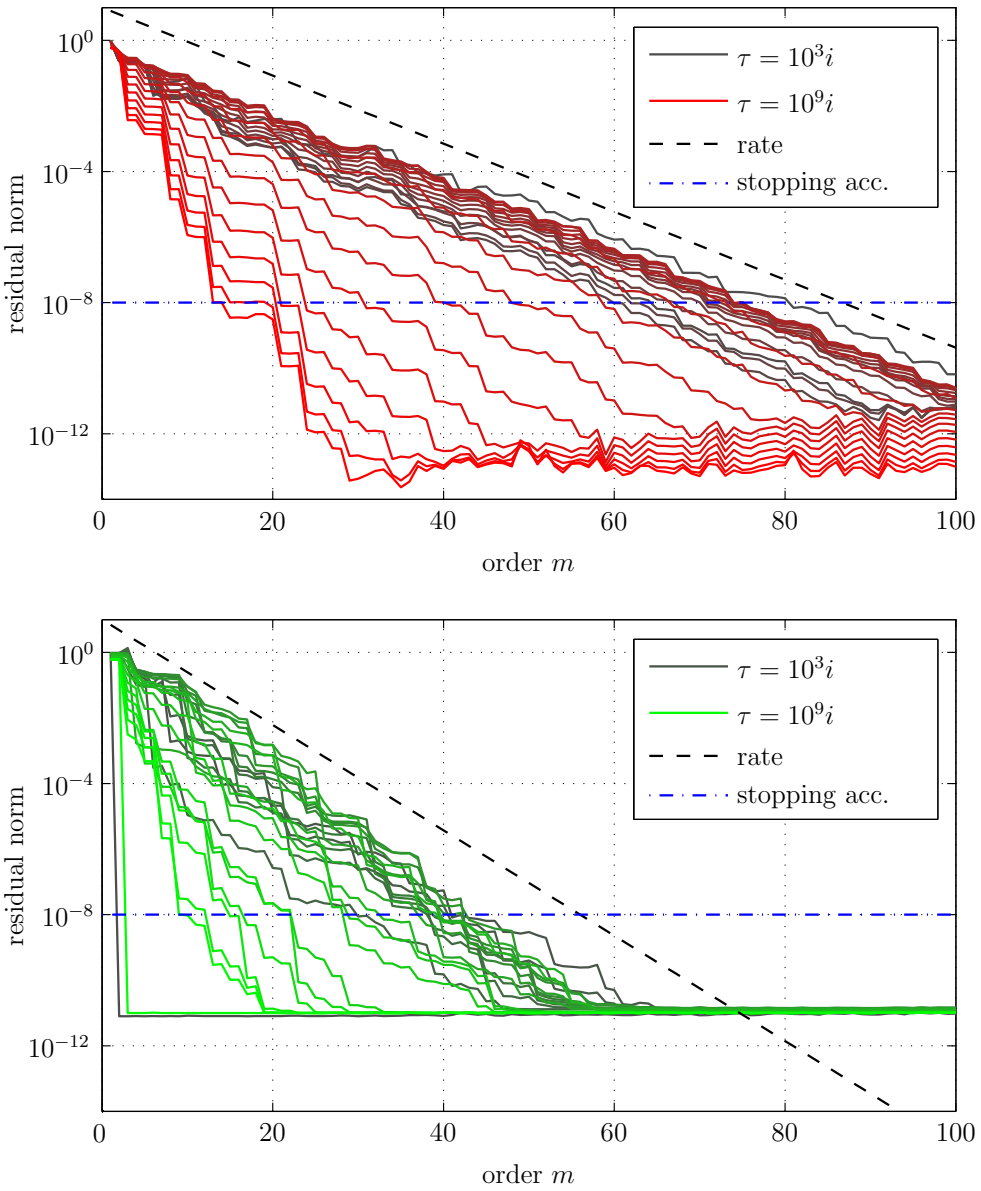


Figure 9.5: Residual curves (solid) and asymptotic convergence rates (dashed) of Rayleigh–Ritz approximations for the resolvent function extracted from a rational Krylov space with real poles (top) and imaginary Zolotarev poles (below).

	Real poles	Zolotarev poles
Iterations m	81	43
PARDISO analysis	1 (1.0 s)	1 (1.0 s)
PARDISO factorizations	13 (50.1 s)	43 (547 s)
PARDISO solves	81 (10.6 s)	43 (9.3 s)
Orthogonalization steps	3321 (3.4 s)	946 (1.7 s)
Compute $\boldsymbol{f}_m^\tau$	25 (2.4 s)	25 (2.1 s)
Total	(67.5 s)	(561.1 s)

Table 9.2: Number of subroutine calls and timings. The problem size is $N = 67937$.

9.2 Lattice Quantum Chromodynamics

The fast computation of the sign function is crucial for simulations in lattice quantum chromodynamics, a physical theory that describes strong interactions between quarks of the constituents of matter [Hig08, §2.7]. A variety of numerical methods have been proposed for this computation and—according to a remark in [EFL$^+$02, p. 5]—all of them turn out to be polynomial Krylov methods. We consider the approximation of $\operatorname{sgn}(Q)\boldsymbol{v}$, where $Q \in \mathbb{C}^{N\times N}$ is the Hermitian form of the Wilson–Dirac operator introduced by Neuberger [Neu98] (see Figure 9.6 for its nonzero structure) and $\boldsymbol{v} \in \mathbb{C}^N$ is a random vector ($N = 3072$). Using the identity $\operatorname{sgn}(z) = z/\sqrt{z^2}$ this problem can be recast as the computation of $f(A)\boldsymbol{b}$ with $f(z) = z^{-1/2}$, $A = Q^2$ and $\boldsymbol{b} = Q\boldsymbol{v}$. Using MATLAB's `eigs` we estimate $\Lambda(A) \subseteq \Sigma = [7.5 \cdot 10^{-4}, 5.1]$. We extract Rayleigh–Ritz approximations from an inexact rational Krylov space, which results from solving the linear systems involved by the CG method with a maximal residual norm of 10^{-10}. As pole sequences we use generalized Leja points on the condenser $(\Sigma, (-\infty, 0])$ and the poles of Zolotarev's best relative approximation for f on Σ of type $(12, 12)$ repeated cyclically. In the former case we expect an asymptotic convergence rate $R = 2.34$ by (7.8), and the Zolotarev poles should be about twice as good (cf. Remark 7.7). To make use of the estimate (6.8) for the sensitivity error it is necessary that the rational Krylov basis V_m is orthogonal and therefore we run the rational Arnoldi algorithm with one reorthogonalization. The error curves (black solid curves) for both pole sequences are shown in Figure 9.7. We observe that the error curve for the method with the Zolotarev poles suddenly drops down to 10^{-8} in iteration $m = 13$, which happens due to the near-optimality property of Rayleigh–Ritz extraction when all 12 poles of the Zolotarev approximation are contained in the rational Krylov space. As expected, about twice as many iterations are required with generalized Leja poles to achieve the same accuracy. We also show the a-posteriori error estimates from Section 6.6, where we have added to all curves the estimate for the sensitivity error from Section 6.3. Therefore the error estimates in Figure 9.7 stagnate, and the stagnation level agrees well with the stagnation of the error curves caused by the inexact linear system solves. Except for the lower bound (6.15), which shows irregular peaks, all error estimators perform reliable.

We remark that the direct evaluation of Zolotarev's rational approximation in partial fraction form by a shifted CG method was advocated in [EFL+02], and the above described "rational" Krylov method is clearly less efficient because the linear systems are solved by an unpreconditioned polynomial Krylov method (see the discussion on page 68). However, taking the detour of first computing a basis of a rational Krylov space and then extracting approximations from it offers the possibility of preconditioning each shifted linear system *independently*, which is (to our best knowledge) not possible if various partial fractions are approximated simultaneously from the same polynomial Krylov space.

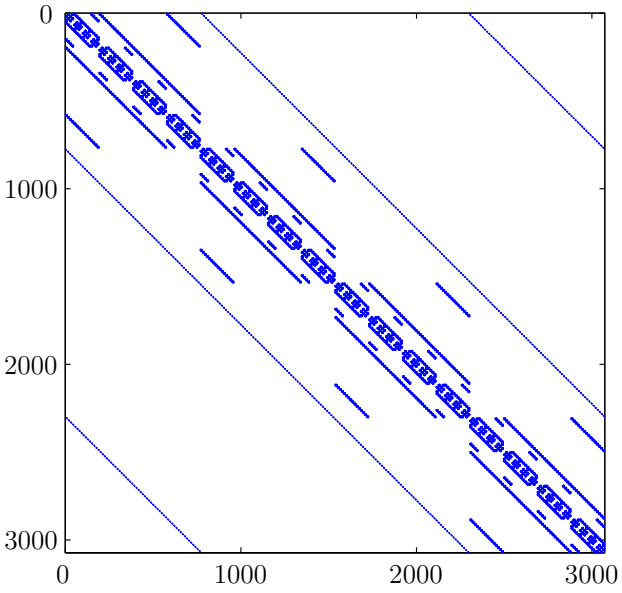


Figure 9.6: Nonzero structure of a Wilson–Dirac matrix Q. This matrix is of size 3072×3072 and has 122880 nonzeros.

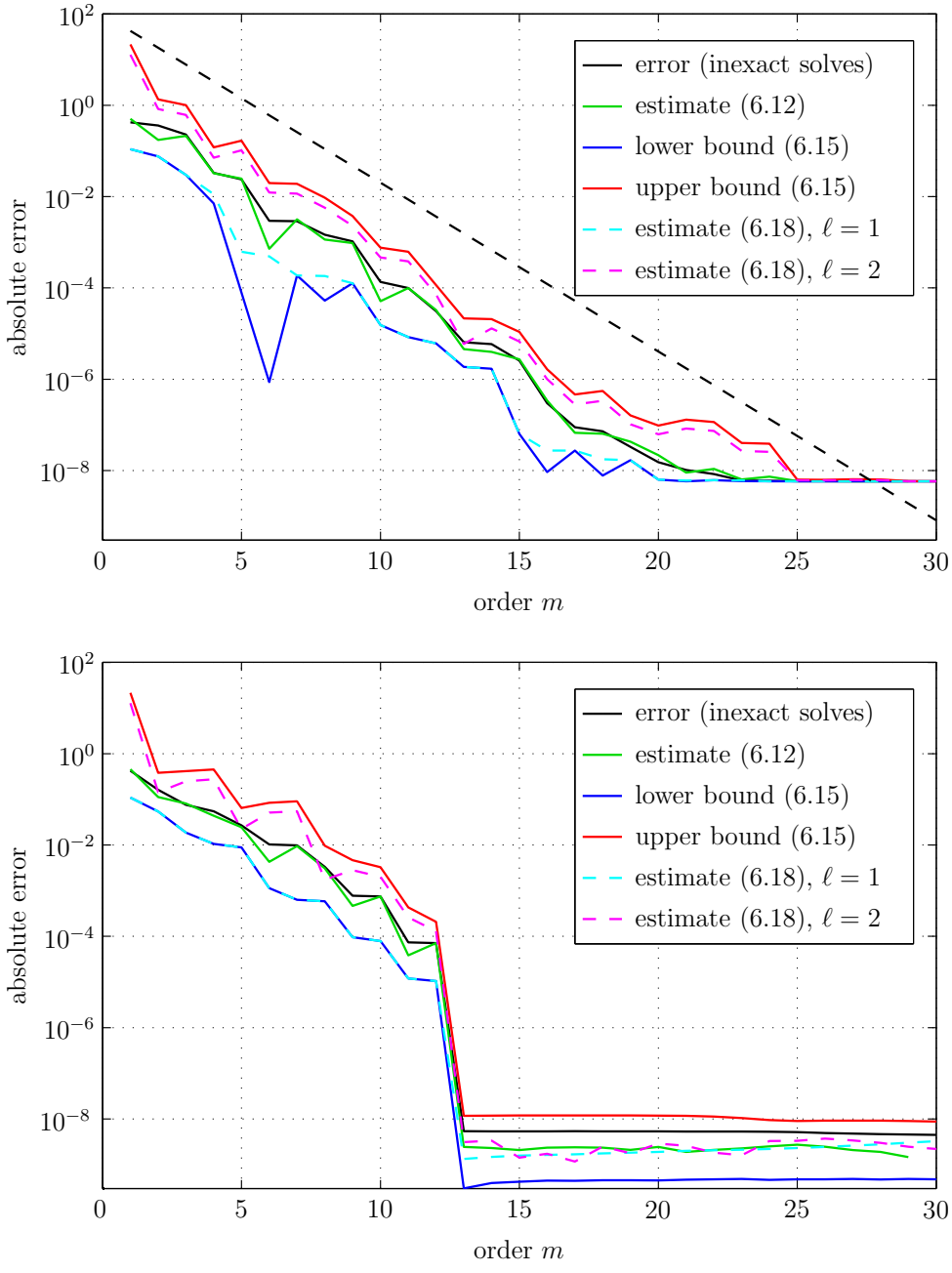


Figure 9.7: Error curves and estimates of inexact Rayleigh–Ritz approximations for the QCD problem $\operatorname{sgn}(Q)\boldsymbol{v}$ using generalized Leja points as poles (top, the asymptotic convergence rate is also shown) and Zolotarev's poles of order 12 (bottom). We added to all estimates for the approximation error the estimated sensitivity error (6.8) in order to predict the level of stagnation caused by the inexact linear system solves.

9.3 An Advection–Diffusion Problem

We consider the initial value problem

$$\begin{aligned}
\partial_\tau u &= \frac{1}{\mathrm{Pe}}\Delta u - \boldsymbol{a}\cdot\nabla u && \text{in } \Omega = (-1,1)\times(0,1),\\
u &= 1-\tanh(\mathrm{Pe}) && \text{on } \Gamma_0,\\
u &= 1+\tanh((2x+1)\,\mathrm{Pe}) && \text{on } \Gamma_{\mathrm{in}},\\
\frac{\partial u}{\partial n} &= 0 && \text{on } \Gamma_{\mathrm{out}},\\
u(\boldsymbol{x},0) &= u_0(\boldsymbol{x}) && \text{in } \Omega
\end{aligned}$$

for the advection–diffusion equation, which is a popular benchmark for discretizations of advection–dominated problems, see [SH82]. The convective field is given as

$$\boldsymbol{a}(x,y) = \begin{bmatrix} 2y(1-x^2) \\ -2x(1-y^2) \end{bmatrix}, \qquad (x,y)\in\Omega,$$

and the boundary $\Gamma = \partial\Omega$ is divided into the inflow boundary $\Gamma_{\mathrm{in}} := [-1,0]\times\{0\}$, the outflow boundary $\Gamma_{\mathrm{out}} := [0,1]\times\{0\}$ and the remaining portion Γ_0, see Figure 9.8 (top). The Péclet number Pe is a nondimensional parameter describing the strength of advection relative to diffusion and therefore also how far the discrete operators are from symmetric. The finite-element discretization of the advection–diffusion operator with $\mathrm{Pe} = 100$ yields a linear ordinary differential equation

$$M\boldsymbol{u}'(\tau) = K\boldsymbol{u}(\tau) + \boldsymbol{g}, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0,$$

with nonsymmetric matrices K, $M \in \mathbb{R}^{N\times N}$ and a constant inhomogeneous term $\boldsymbol{g}\in\mathbb{R}^N$ resulting from the inhomogeneous Dirichlet boundary condition ($N = 2912$). We then approximate the matrix exponential part of the solution

$$\boldsymbol{u}(\tau) = \exp(\tau A)(\boldsymbol{u}_0 + K^{-1}\boldsymbol{g}) - K^{-1}\boldsymbol{g}, \quad A = M^{-1}K$$

at time $\tau = 1$, starting with the initial value $\boldsymbol{u}(0) = \boldsymbol{u}_0 = \boldsymbol{0}$, from a rational Krylov space with all poles $\xi_j = 10$. To simulate the effect of inexact solves we use the GMRES method to solve the linear systems involved with a maximal residual norm of 10^{-10}.

In Figure 9.8 (bottom) we show the spectrum $\Lambda(A)$ and parts of the numerical range $\mathbb{W}(A)$. For the error bound (6.16) we would have to search the set $\mathbb{W}(A)$ for a maximum of the function $g_m(\zeta)$ defined in (6.14), but this is surely not practical. Instead we only search the interval $[\lambda_{\min}, \lambda_{\max}]$ spanned by the smallest and largest real eigenvalue of A. This no longer guarantees an error bound, but it can still serve as an error estimate, see Figure 9.9. Again we have added to all error estimates and bounds the estimate for the sensitivity error from Section 6.3. Note that the level of stagnation due to the inexact solves is predicted well, even though A is highly nonnormal.

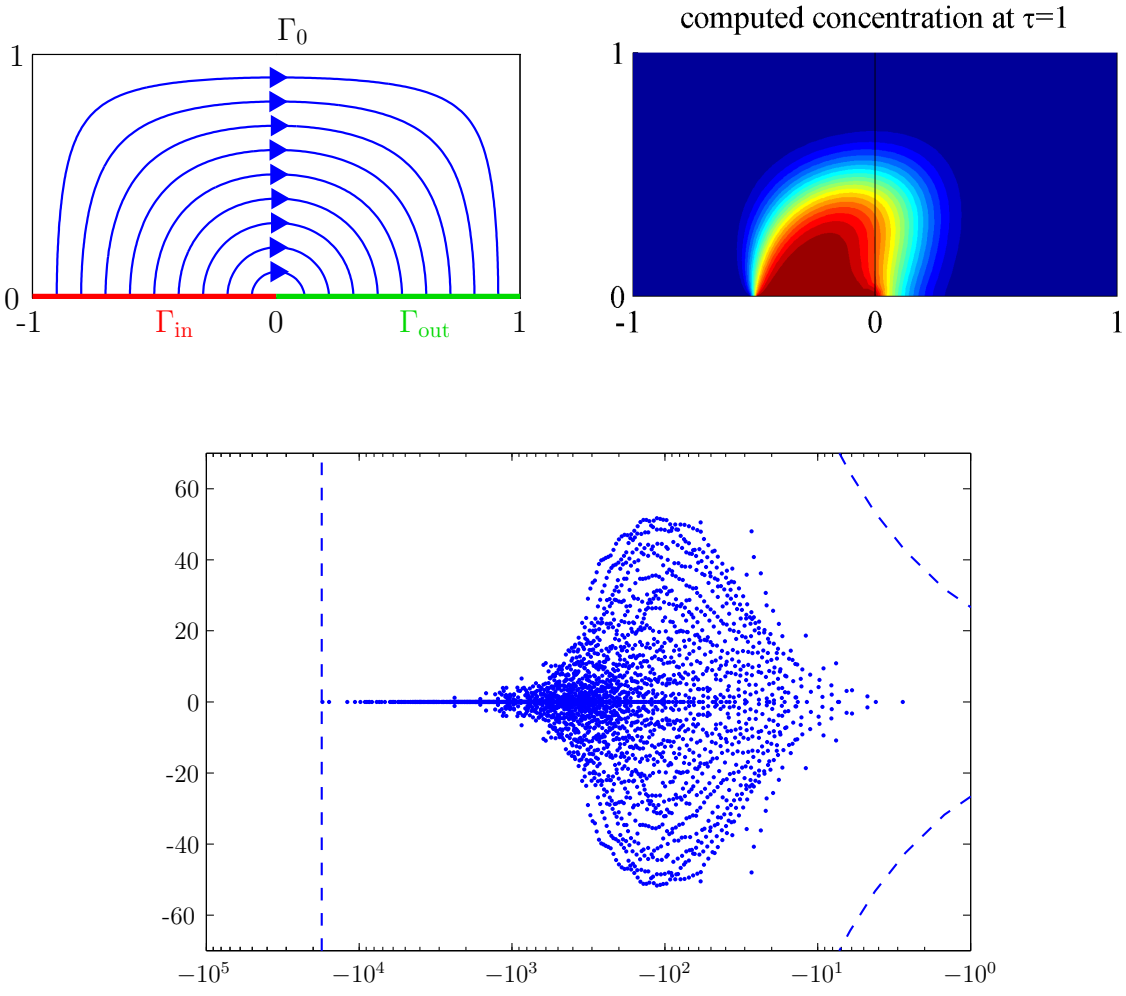


Figure 9.8: Spatial domain and a solution of the advection–diffusion problem. Below is shown the spectrum of A and parts of its numerical range.

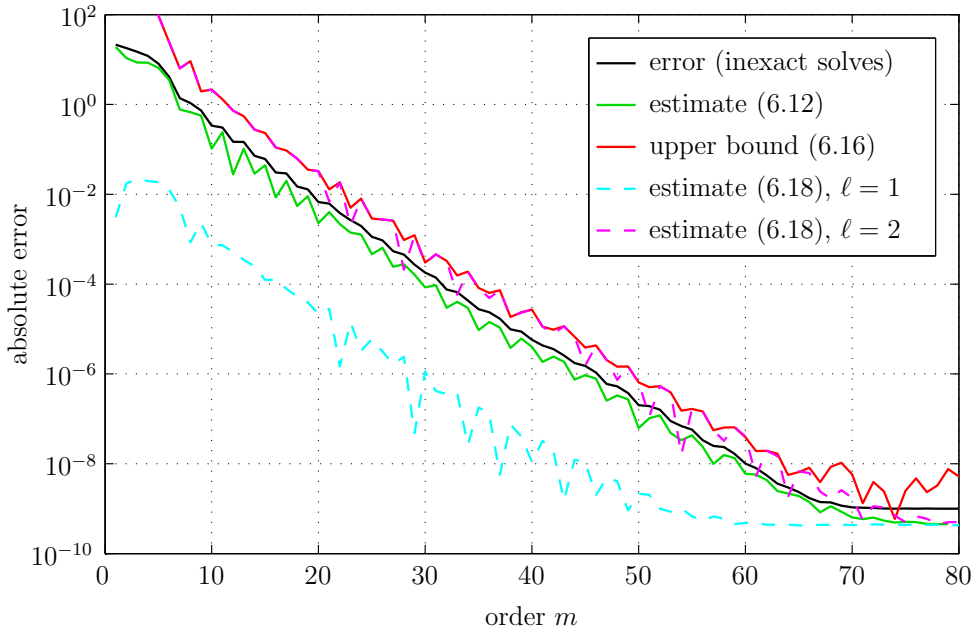


Figure 9.9: Error curves and estimates of inexact Rayleigh–Ritz approximations for the advection–diffusion problem. We added to all estimates for the approximation error the estimated sensitivity error (6.8).

9.4 A Wave Equation

For the differential operator $A = -\partial_{xx}$ the problem

$$\begin{aligned} \partial_{\tau\tau} u &= -Au && \text{in } \Omega = (0,\pi),\ \tau > 0, \\ u(0,\tau) = u(\pi,\tau) &= 0 && \text{for all } \tau \geq 0, \\ u(x,0) &= u_0(x) && \text{for all } x \in \Omega, \\ \partial_\tau u(x,0) &= u_1(x) && \text{for all } x \in \Omega, \end{aligned}$$

has the solution

$$u(x,\tau) = \cos(\tau A^{1/2})u_0(x) + \tau\,\mathrm{sinc}(\tau A^{1/2})u_1(x),$$

where $\mathrm{sinc}(x) = \sin(x)/x$, provided that the initial data is smooth enough, i.e., $u_0 \in \mathcal{D}(A)$ and $u_1 \in \mathcal{D}(A^{1/2})$ (cf. [GH08]). The eigenvalues and eigenvectors of A respecting the homogenous Dirichlet boundary conditions are

$$\lambda_j = j^2, \quad \hat{u}_j(x) = \sqrt{2/\pi}\sin(jx) \qquad (j = 1, 2, \ldots).$$

To obtain a simple example we let

$$u(x,\tau) = e^{\cos(\tau+x)} - e^{\cos(\tau-x)},$$

which is obviously a solution of the above differential equation with $u(x,0) = u_0(x) \equiv 0$. We therefore have

$$u(x,\tau) = f^\tau(A)\boldsymbol{b}, \quad \text{where } f^\tau(z) = \tau\,\mathrm{sinc}(\tau z^{1/2}),\ \boldsymbol{b} = \partial_\tau u(x,0) = u_1.$$

The chebfun system [THP+09, PPT10] allows us, in conjunction with the chebop functionality [DBT08], to do MATLAB computations with the unbounded operator A (based on spectral collocation methods on Chebyshev grids of automatically determined resolution). In particular, we can easily implement a rational Arnoldi algorithm by first defining the operators A and I with the commands

```
omega = domain(0,pi);
A = -diff(omega,2);
I = eye(omega);
```

A typical rational Arnoldi step then reads as

```
x = (I-A/xi(j) & 'dirichlet') \ (A*y);
```

where `x` and `y` are continuous functions (chebfuns) on the interval $[0, \pi]$, followed by the orthogonalization of `x` with respect to the inner product `y'*x` of $L^2([0, \pi])$. In Figure 9.10 we show the error of Rayleigh–Ritz approximations of order $m = 3$ computed for 5 different parameters $\tau = 2j\pi/5$ $(j = 1, \ldots, 5)$. All poles of the rational Krylov space are chosen rather arbitrary at $\xi = -10$. For $m = 4$ the Rayleigh–Ritz approximations become already visually indistinguishable from the exact solutions $u(x, \tau)$ (shown as dashed lines) and for $m = 10$ the L^2-error is less than 10^{-8}. In Figure 9.11 the rational Ritz values of orders $m = 1, \ldots, 100$ are plotted in different colors indicating the distance to a closest eigenvalue of A (cf. Table 8.1 on page 127). As expected, the rational Ritz values start converging near the left endpoint of the spectrum close to the pole ξ.

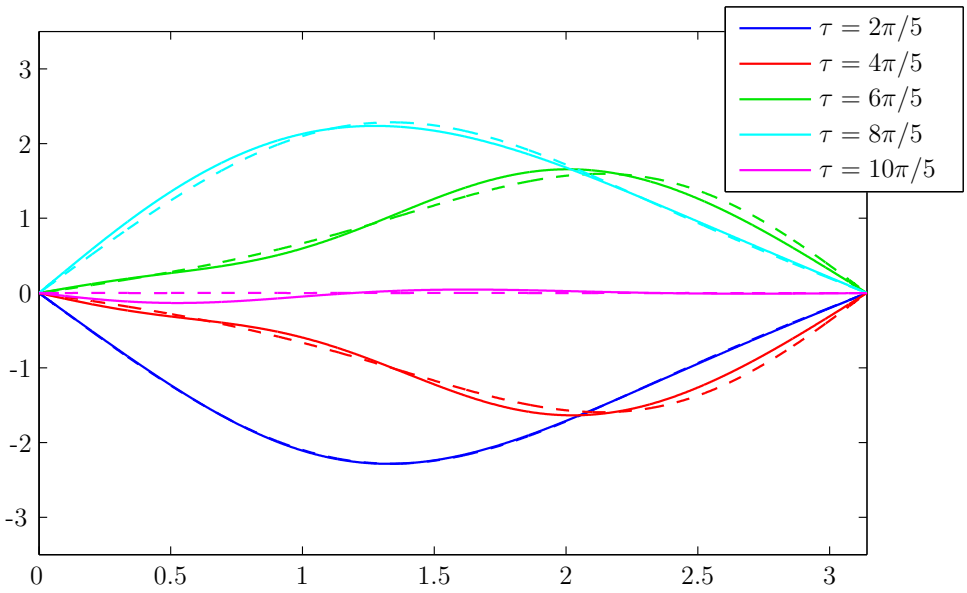


Figure 9.10: Rayleigh–Ritz approximations of order $m = 3$ for the 1D wave equation evaluated for different time parameters τ (solid lines) and the exact solutions $u(x, \tau)$ (dashed lines).

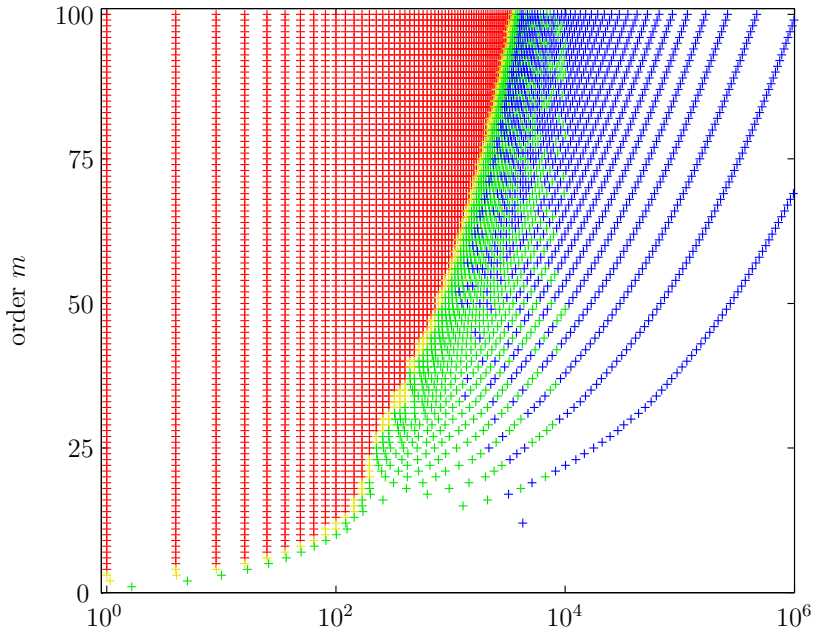


Figure 9.11: Rational Ritz values for the differential operator $A = -\partial_{xx}$ with homogeneous Dirichlet boundary conditions. All poles of the rational Krylov space are at $\xi = -10$.

Bibliography

[ABB00] E. J. Allen, J. Baglama, and S. K. Boyd. Numerical approximation of the product of the square root of a matrix with a vector. *Linear Algebra Appl.*, 310:167–181, 2000.

[AEEG08a] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel. A generalization of the steepest descent method for matrix functions. *Electron. Trans. Numer. Anal.*, 28:206–222, 2008.

[AEEG08b] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel. Implementation of a restarted Krylov subspace method for the evaluation of matrix functions. *Linear Algebra Appl.*, 429:2293–2314, 2008.

[Ahl73] L. V. Ahlfors. *Conformal Invariants: Topics in Geometric Function Theory.* McGraw-Hill, New York, 1973.

[AK00] O. Axelsson and A. Kucherov. Real valued iterative methods for solving complex symmetric linear systems. *Numer. Linear Algebra Appl.*, 7:197–218, 2000.

[AL08] M. Arioli and D. Loghin. Matrix square-root preconditioners for the Steklov–Poincaré operator. Technical Report RAL-TR-2008-003, Rutherford Appleton Laboratory, Didcot, UK, 2008.

[All99] E. J. Allen. Stochastic differential equations and persistence time for two interacting populations. *Dyn. Contin. Discrete Impuls. Systems*, 5:271–281, 1999.

[And81] J.-E. Andersson. Approximation of e^{-x} by rational functions with concentrated negative poles. *J. Approx. Theory*, 32:85–95, 1981.

[Arn51] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[AS84] M. Abramowitz and I. A. Stegun. *Pocketbook of Mathematical Functions.* Verlag Harri Deutsch, Thun, 1984.

[Bag67] T. Bagby. The modulus of a plane condenser. *J. Math. Mech.*, 17:315–329, 1967.

[Bag69] T. Bagby. On interpolation by rational functions. *Duke Math. J.*, 36:95–104, 1969.

[Bai02] Z. Bai. Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Appl. Numer. Math.*, 43:9–44, 2002.

[BB03] M. Benzi and D. Bertaccini. Approximate inverse preconditioning for shifted linear systems. *BIT*, 43:231–244, 2003.

[BC07] B. Beckermann and S. S. Capizzano. On the asymptotic spectrum of finite element matrix sequences. *SIAM J. Numer. Anal.*, 45:746–769, 2007.

[BCV03] L. Bergamaschi, M. Caliari, and M. Vianello. Efficient approximation of the exponential operator for discrete 2D advection–diffusion problems. *Numer. Linear Algebra Appl.*, 10:271–289, 2003.

[BCV04] L. Bergamaschi, M. Caliari, and M. Vianello. The ReLPM exponential integrator for FE discretizations of advection–diffusion equations. In M. Bubak et al., editors, *Computational Science – ICCS 2004*, volume 3039 of *Lecture Notes in Computer Science*, pages 434–442. Springer-Verlag, Berlin, 2004.

[Bec00a] B. Beckermann. A note on the convergence of Ritz values for sequences of matrices. Technical Report ANO 408, Université de Lille I, Labo Paul Painlevé, Lille, France, 2000.

[Bec00b] B. Beckermann. On a conjecture of E. A. Rakhmanov. *Constr. Approx.*, 16:427–448, 2000.

[Bec06] B. Beckermann. Discrete orthogonal polynomials and superlinear convergence of Krylov subspace methods in numerical linear algebra. In F. Marcellan and W. Van Assche, editors, *Orthogonal Polynomials and Special Functions*, volume 1883 of *Lecture Notes in Mathematics*, pages 119–185. Springer-Verlag, Berlin, 2006.

[Bel70] R. Bellman. *Introduction to Matrix Analysis.* McGraw-Hill, New York, 2nd edition, 1970.

[Ber04] D. Bertaccini. Efficient preconditioning for sequences of parametric complex symmetric linear systems. *Electron. Trans. Numer. Anal.*, 18:49–64, 2004.

[BES08] R. Börner, O. G. Ernst, and K. Spitzer. Fast 3D simulation of transient electromagnetic fields by model reduction in the frequency domain using Krylov subspace projection. *Geophys. J. Int.*, 173:766–780, 2008.

[Bet08] T. Betcke. Optimal scaling of generalized and polynomial eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 30:1320–1338, 2008.

[BG03] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses. Theory and Applications.* Springer-Verlag, New York, 2nd edition, 2003.

[BG10] B. Beckermann and A. Gryson. Extremal rational functions on symmetric discrete sets and superlinear convergence of the ADI method. To appear in Constr. Approx., 2010.

[BGHN99] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad. *Orthogonal Rational Functions.* Cambridge University Press, Cambridge, UK, 1999.

[BGHN03] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad. Orthogonal rational functions and tridiagonal matrices. *J. Comput. Appl. Math.*, 153:89–97, 2003.

[BGV10] B. Beckermann, S. Güttel, and R. Vandebril. On the convergence of rational Ritz values. *SIAM J. Matrix Anal. Appl.*, 31:1740–1774, 2010.

[Bjö67] A. Björck. Solving linear least squares problems by Gram–Schmidt orthogonalization. *BIT*, 7:1–21, 1967.

[BK01a] B. Beckermann and A. B. J. Kuijlaars. On the sharpness of an asymptotic error estimate for conjugate gradients. *BIT*, 41:856–867, 2001.

[BK01b] B. Beckermann and A. B. J. Kuijlaars. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal.*, 39:300–329, 2001.

[BK02] B. Beckermann and A. B. J. Kuijlaars. Superlinear CG convergence for special right-hand sides. *Electron. Trans. Numer. Anal.*, 14:1–19, 2002.

[Boo91] C. de Boor. An alternative approach to (the teaching of) rank, basis, and dimension. *Linear Algebra Appl.*, 146:221–229, 1991.

[Bor83] P. B. Borwein. Rational approximations with real poles to e^{-x} and x^n. *J. Approx. Theory*, 38:279–283, 1983.

[Bot02] M. Botchev. A. N. Krylov: a short biography. `http://www.math.uu.nl/people/vorst/kryl.html`, 2002.

[BP92] Å. Björck and C. C. Paige. Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm. *SIAM J. Matrix Anal. Appl.*, 13:176–190, 1992.

[BR99] V. Buyarov and E. A. Rakhmanov. Families of equilibrium measures with external field on the real axis. *Sb. Math.*, 190:791–802, 1999.

[BR09] B. Beckermann and L. Reichel. Error estimation and evaluation of matrix functions via the Faber transform. *SIAM J. Numer. Anal.*, 47:3849–3883, 2009.

[Bri87] W. L. Briggs. *A Multigrid Tutorial.* SIAM, Philadelphia, PA, 1987.

[BS99] A. Böttcher and B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices.* Springer-Verlag, New York, 1999.

[BT04] Z. Battles and L. N. Trefethen. An extension of MATLAB to continuous functions and operators. *SIAM J. Sci. Comput.*, 25:1743–1770, 2004.

[Buc84] A. Buchheim. On the theory of matrices. *Proc. London Math. Soc.*, 16:63–82, 1884.

[Buc86] A. Buchheim. An extension of a theorem of Professor Sylvester's relating to matrices. *Phil. Mag.*, 22:173–174, 1886. Fifth series.

[BV00] L. Bergamaschi and M. Vianello. Efficient computation of the exponential operator for large, sparse, symmetric matrices. *Numer. Linear Algebra Appl.*, 7:27–45, 2000.

[Cal07] M. Caliari. Accurate evaluation of divided differences for polynomial interpolation of exponential propagators. *Computing*, 80:189–201, 2007.

[Cay58] A. Cayley. A memoir on the theory of matrices. *Philos. Trans. Roy. Soc. London*, 148:17–37, 1858.

[Cay72] A. Cayley. On the extraction of the square root of a matrice of the third order. *Proc. Roy. Soc. Edinburgh*, 7:675–682, 1872.

[CGR95] D. Calvetti, E. Gallopoulos, and L. Reichel. Incomplete partial fractions for parallel evaluation of rational matrix functions. *J. Comput. Appl. Math.*, 59:349–380, 1995.

[Che98] E. W. Cheney. *Introduction to Approximation Theory.* Chelsea Publishing, New York, 1998. Reprint of the 2nd edition from 1982.

[CM02] S. M. Cox and P. C. Matthews. Exponential time differencing for stiff systems. *J. Comput. Phys.*, 176:430–455, 2002.

[CMV69] W. J. Cody, G. Meinardus, and R. S. Varga. Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems. *J. Approx. Theory*, 2:50–65, 1969.

[Col89] J. P. Coleman. Numerical methods for $y'' = f(x, y)$ via rational approximations for the cosine. *IMA J. Numer. Anal.*, 9:145–165, 1989.

[Cro07] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *J. Funct. Anal.*, 244:668–690, 2007.

[CRZ99] D. Calvetti, L. Reichel, and Q. Zhang. Iterative exponential filtering for large discrete ill-posed problems. *Numer. Math.*, 83:535–556, 1999.

[CV05] J. Coussement and W. Van Assche. A continuum limit of relativistic Toda lattice: asymptotic theory of discrete Laurent orthogonal polynomials with varying recurrence coefficients. *J. Phys. A*, 38:3337–3366, 2005.

[CVB04] M. Caliari, M. Vianello, and L. Bergamaschi. Interpolating discrete advection–diffusion propagators at Leja sequences. *J. Comput. Appl. Math.*, 172:79–99, 2004.

[CVB07] M. Caliari, M. Vianello, and L. Bergamaschi. The LEM exponential integrator for advection–diffusion–reaction equations. *J. Comput. Appl. Math.*, 210:56–63, 2007.

[Däp88] H. D. Däppen. *The Schwarz–Christoffel Map for Doubly Connected Domains with Applications.* PhD thesis, ETH Zürich, Zürich, Switzerland, 1988. In German.

[Dav59] P. J. Davis. On the numerical integration of periodic analytic functions. In E. R. Langer, editor, *On Numerical Approximation*, pages 21–23. University of Wisconsin Press, Madison, WI, 1959.

[DB07] K. Deckers and A. Bultheel. Rational Krylov sequences and orthogonal rational functions. Technical Report TW499, Katholieke Universiteit Leuven, Department of Computer Science, Leuven, Belgium, 2007.

[DBT08] T. A. Driscoll, F. Bornemann, and L. N. Trefethen. The chebop system for automatic solution of differential equations. *BIT*, 48:701–723, 2008.

[DDSV98] J. J. Dongarra, I. S. Duff, D. C. Sorensen, and H. A. van der Vorst. *Numerical Linear Algebra for High-Performance Computers.* SIAM, Philadelphia, PA, 1998.

[DES82] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Anal.*, 19:400–408, 1982.

[DK89] V. L. Druskin and L. A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *USSR Comput. Maths. Math. Phys.*, 29:112–121, 1989.

[DK94] V. L. Druskin and L. A. Knizhnerman. Spectral approach to solving three-dimensional Maxwell's diffusion equations in the time and frequency domains. *Radio Science*, 29:937–953, 1994.

[DK98] V. L. Druskin and L. A. Knizhnerman. Extended Krylov subspaces: Approximation of the matrix square root and related functions. *SIAM J. Matrix Anal. Appl.*, 19:775–771, 1998.

[DKZ09] V. L. Druskin, L. A. Knizhnerman, and M. Zaslavsky. Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts. *SIAM J. Sci. Comput.*, 31:3760–3780, 2009.

[DMR08] F. Diele, I. Moret, and S. Ragni. Error estimates for polynomial Krylov approximations to matrix functions. *SIAM J. Matrix Anal. Appl.*, 30:1546–1565, 2008.

[DR84] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration.* Academic Press, New York, 2nd edition, 1984.

[Dri96] T. A. Driscoll. Algorithm 756: A MATLAB toolbox for Schwarz–Christoffel mapping. *ACM Trans. Math. Software*, 22:168–186, 1996.

[Dri05] T. A. Driscoll. Algorithm 843: Improvements to the Schwarz–Christoffel toolbox for MATLAB. *ACM Trans. Math. Software*, 31:239–251, 2005.

[DS58] N. Dunford and J. T. Schwartz. *Linear Operators. Part I: General Theory.* Interscience, New York, 1958.

[DS80] V. A. Dougalis and S. M. Serbin. Some remarks on a class of rational approximations to the cosine. *BIT*, 20:204–211, 1980.

[DS97] P. D. Dragnev and E. B. Saff. Constrained energy problems with applications to orthogonal polynomials of a discrete variable. *J. Anal. Math.*, 72:223–259, 1997.

[DTT98] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen. From potential theory to matrix iterations in six steps. *SIAM Rev.*, 40:547–578, 1998.

[Duf97] I. S. Duff. Sparse numerical linear algebra: Direct methods and preconditioning. In I. S. Duff and G. A. Watson, editors, *The State of the Art in Numerical Analysis*, pages 27–62. Oxford University Press, Oxford, UK, 1997.

[Dun43] N. Dunford. Spectral theory. *Trans. Amer. Math. Soc.*, 54:185–217, 1943.

[EE06] M. Eiermann and O. G. Ernst. A restarted Krylov subspace method for the evaluation of matrix functions. *SIAM J. Numer. Anal.*, 44:2481–2504, 2006.

[EEG09] M. Eiermann, O. G. Ernst, and S. Güttel. Deflated restarting for matrix functions. Submitted to SIAM J. Matrix Anal. Appl., 2009.

[EFL+02] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. A. van der Vorst. Numerical methods for the QCD overlap operator. I: Sign-function and error bounds. *Comput. Phys. Commun.*, 146:203–224, 2002.

[EH06] J. van den Eshof and M. Hochbruck. Preconditioning Lanczos approximations to the matrix exponential. *SIAM J. Sci. Comput.*, 27:1438–1457, 2006.

[Eie84] M. Eiermann. On the convergence of Padé-type approximants to analytic functions. *J. Comput. Appl. Math.*, 10:219–227, 1984.

[Ell83] S. W. Ellacott. On the Faber transform and efficient numerical rational approximation. *SIAM J. Numer. Anal.*, 20:989–1000, 1983.

[ER80] T. Ericsson and A. Ruhe. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Math. Comp.*, 35:1251–1268, 1980.

[ER82] T. Ericsson and A. Ruhe. STLM – a software package for the spectral transformation Lanczos algorithm. Technical Report UMINF-101, Umeå University, Department of Computer Science, Umeå, Sweden, 1982.

[Eri90] T. Ericsson. Computing functions of matrices using Krylov subspace methods. Technical Report, Chalmers University of Technology, Department of Computer Science, Göteborg, Sweden, 1990.

[ES04] J. van den Eshof and G. L. G. Sleijpen. Inexact Krylov subspace methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 26:125–153, 2004.

[ESG05] J. van den Eshof, G. L. G. Sleijpen, and M. B. van Gijzen. Relaxation strategies for nested Krylov methods. *J. Comput. Appl. Math.*, 177:347–365, 2005.

[EW91] N. S. Ellner and E. L. Wachspress. Alternating direction implicit iteration for systems with complex spectra. *SIAM J. Numer. Anal.*, 28:859–870, 1991.

[Fas05] D. Fasino. Rational Krylov matrices and QR steps on Hermitian diagonal-plus-semiseparable matrices. *Numer. Linear Algebra Appl.*, 12:743–754, 2005.

[FF76] D. K. Faddejew and W. N. Faddejewa. *Numerische Methoden der linearen Algebra.* Oldenbourg Verlag, München, 1976.

[FG98] A. Frommer and U. Glässner. Restarted GMRES for shifted linear systems. *SIAM J. Sci. Comput.*, 19:15–26, 1998.

[Fre90] R. W. Freund. On conjugate gradient type methods and polynomial preconditioners for a class of complex non-Hermitian matrices. *Numer. Math.*, 57:285–312, 1990.

[Fre93] R. W. Freund. Solution of shifted linear systems by quasi-minimal residual iterations. In L. Reichel et al., editors, *Numerical Linear Algebra*, pages 101–121. Walter de Gruyter, Berlin, Berlin, 1993.

[Fre03] R. W. Freund. Model reduction methods based on Krylov subspaces. *Acta Numer.*, 12:267–319, 2003.

[Fro96] G. Frobenius. Ueber die cogredienten Transformationen der bilinearen Formen. *Sitzungsber. K. Preuss. Akad. Wiss. Berlin*, 16:7–16, 1896.

[Fro03] A. Frommer. BiCGStab(ℓ) for families of shifted linear systems. *Computing*, 70:87–109, 2003.

[FS92] A. Frommer and D. B. Szyld. H-splittings and two-stage iterative methods. *Numer. Math.*, 63:345–356, 1992.

[FS08a] A. Frommer and V. Simoncini. Matrix functions. In W. H. A. Schilders et al., editors, *Model Order Reduction: Theory, Research Aspects and Applications*, pages 275–303. Springer-Verlag, Berlin, 2008.

[FS08b] A. Frommer and V. Simoncini. Stopping criteria for rational matrix functions of Hermitian and symmetric matrices. *SIAM J. Sci. Comput.*, 30:1387–1412, 2008.

[FTDR89] R. A. Friesner, L. S. Tuckerman, B. C. Dornblaser, and T. V. Russo. A method for exponential propagation of large systems of stiff nonlinear differential equations. *J. Sci. Comput.*, 4:327–354, 1989.

[Gan59] F. R. Gantmacher. *The Theory of Matrices*, volume one. Chelsea Publishing, New York, 1959.

[GGV96] K. Gallivan, E. Grimme, and P. Van Dooren. A rational Lanczos algorithm for model reduction. *Numer. Algorithms*, 12:33–63, 1996.

[GH08] V. Grimm and M. Hochbruck. Rational approximation to trigonometric operators. *BIT*, 48:215–229, 2008.

[GHK02] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. $\mathcal{H}$-matrix approximation for the operator exponential with applications. *Numer. Math.*, 92:83–111, 2002.

[GHK03] L. Grasedyck, W. Hackbusch, and B. N. Khoromskij. Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices. *Computing*, 70:121–165, 2003.

[GHK04] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. Data-sparse approximation to the operator-valued functions of elliptic operator. *Math. Comp.*, 73:1297–1324, 2004.

[GHK05] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. Data-sparse approximation to a class of operator-valued functions. *Math. Comp.*, 74:681–708, 2005.

[GHNS86] Y. Goldman, C. Hubans, S. Nicoletis, and S. Spitz. A finite-element solution for the transient electromagnetic response of an arbitrary two-dimensional resistivity distribution. *Geophysics*, 51:1450–1461, 1986.

[GLR02] L. Giraud, J. Langou, and M. Rozložník. On the round-off error analysis of the Gram–Schmidt algorithm with reorthogonalization. Research Report TR/PA/02/33, CERFACS, Toulouse, France, 2002.

[GLR05] L. Giraud, J. Langou, and M. Rozložník. The loss of orthogonality in the Gram–Schmidt orthogonalization process. *Comput. Math. Appl.*, 50:1069–1075, 2005.

[Gon69] A. A. Gonchar. Zolotarev problems connected with rational functions. *Math. USSR Sb.*, 7:623–635, 1969.

[Gon78] A. A. Gonchar. On the speed of rational approximation of some analytic functions. *Math. USSR Sb.*, 34:131–145, 1978.

[GRS97] A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical behaviour of the modified Gram–Schmidt GMRES implementation. *BIT*, 37:706–719, 1997.

[GS89] E. Gallopoulos and Y. Saad. On the parallel solution of parabolic equations. In R. de Groot, editor, *Proceedings of the International Conference on Supercomputing 1989*, pages 17–28. ACM Press, New York, 1989.

[GS92] E. Gallopoulos and Y. Saad. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM J. Sci. Statist. Comput.*, 13:1236–1264, 1992.

[GSH07] N. I. M. Gould, J. A. Scott, and Y. Hu. A numerical evaluation of sparse direct solvers for the solution of large sparse symmetric linear systems of equations. *ACM Trans. Math. Software*, 33:10, 2007.

[GT83] M. H. Gutknecht and L. N. Trefethen. Real and complex Chebyshev approximation on the unit disk and interval. *Bull. Amer. Math. Soc.*, 8:455–458, 1983.

[Gup02] A. Gupta. Recent advances in direct methods for solving unsymmetric sparse systems of linear equations. *ACM Trans. Math. Software*, 28:301–324, 2002.

[GV96] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[GY00] G. H. Golub and Q. Ye. Inexact inverse iteration for generalized eigenvalue problems. *BIT*, 40:671–684, 2000.

[Hal72] P. R. Halmos. *Introduction to Hilbert Space and the Theory of Spectral Multiplicity.* Chelsea Publishing, New York, 2nd edition, 1972.

[Hal82] P. R. Halmos. *A Hilbert Space Problem Book.* Springer-Verlag, New York, 2nd edition, 1982.

[Hen71] P. Henrici. An algorithm for the incomplete decomposition of a rational function into partial fractions. *Z. Angew. Math. Phys.*, 22:751–755, 1971.

[Hen88] P. Henrici. *Applied and Computational Complex Analysis*, volume I. John Wiley & Sons, New York, 1988.

[Hen93] P. Henrici. *Applied and Computational Complex Analysis*, volume III. John Wiley & Sons, New York, 1993.

[HGB02] W. Hackbusch, L. Grasedyck, and S. Börm. An introduction to hierarchical matrices. *Math. Bohem.*, 127:229–241, 2002.

[HH05] M. Hochbruck and M. E. Hochstenbach. Subspace extraction for matrix functions. Preprint, Case Western Reserve University, Department of Mathematics, Cleveland, OH, 2005.

[HHT08] N. Hale, N. J. Higham, and L. N. Trefethen. Computing A^α, $\log(A)$ and related matrix functions by contour integrals. *SIAM J. Numer. Anal.*, 46:2505–2523, 2008.

[Hig02] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia, PA, 2nd edition, 2002.

[Hig08] N. J. Higham. *Functions of Matrices. Theory and Computation.* SIAM, Philadelphia, PA, 2008.

[HKV05] S. Helsen, A. B. J. Kuijlaars, and M. Van Barel. Convergence of the isometric Arnoldi process. *SIAM J. Matrix Anal. Appl.*, 26:782–809, 2005.

[HL97] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34:1911–1925, 1997.

[HLS98] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Statist. Comput.*, 19:1552–1574, 1998.

[HPKS99] W. Huisinga, L. Pesce, R. Kosloff, and P. Saalfrank. Faber and Newton polynomial integrators for open-system density matrix propagation. *J. Chem. Phys.*, 110:5538–5547, 1999.

[HT88] L. Halpern and L. N. Trefethen. Wide-angle one-way wave equations. *J. Acoust. Soc. Amer.*, 84:1397–1404, 1988.

[ITS10] M. Ilić, I. W. Turner, and D. P. Simpson. A restarted Lanczos approximation to functions of a symmetric matrix. To appear in IMA J. Numer. Anal., 2010.

[JR09] C. Jagels and L. Reichel. The extended Krylov subspace method and orthogonal Laurent polynomials. *Linear Algebra Appl.*, 431:441–458, 2009.

[KDZ09] L. A. Knizhnerman, V. L. Druskin, and M. Zaslavsky. On optimal convergence rate of the rational Krylov subspace reduction for electromagnetic problems in unbounded domains. *SIAM J. Numer. Anal.*, 47:953–971, 2009.

[Ken04] A. D. Kennedy. Approximation theory for matrices. *Nucl. Phys. B*, 128:107–116, 2004.

[KGGK94] V. Kumar, A. Grama, A. Gupta, and G. Karypis. *Introduction to Parallel Computing. Design and Analysis of Algorithms.* Benjamin/Cummings Publishing, Redwood City, CA, 1994.

[KMS53] M. Kac, W. L. Murdock, and G. Szegő. On the eigenvalues of certain Hermitian forms. *Indiana Univ. Math. J.* (formerly known as *Journal of Rational Mechanics and Analysis*), 2:767–800, 1953.

[Kni91] L. A. Knizhnerman. Calculation of functions of unsymmetric matrices using Arnoldi's method. *USSR Comput. Maths. Math. Phys.*, 31:1–9, 1991.

[KR98] A. B. J. Kuijlaars and E. A. Rakhmanov. Zero distributions for discrete orthogonal polynomials. *J. Comput. Appl. Math.*, 99:255–274, 1998.

[Kra99] S. G. Krantz. *Handbook of Complex Variables.* Birkhäuser, Boston, MA, 1999.

[Kre99] R. Kress. *Linear Integral Equations.* Springer-Verlag, Berlin, second edition, 1999.

[Kry31] A. N. Krylov. On the numerical solution of the equation by which, in technical matters, frequencies of small oscillations of material systems are determined. *Izv. Akad. Nauk SSSR*, 1:555–570, 1931. In Russian.

[KS10] L. A. Knizhnerman and V. Simoncini. A new investigation of the extended Krylov subspace method for matrix function evaluations. To appear in Numer. Linear Algebra Appl., 2010.

[KT05] A.-K. Kassam and L. N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM J. Sci. Comput.*, 26:1214–1233, 2005.

[Kui00] A. B. J. Kuijlaars. Which eigenvalues are found by the Lanczos method? *SIAM J. Matrix Anal. Appl.*, 22:306–321, 2000.

[Kui06] A. B. J. Kuijlaars. Convergence analysis of Krylov subspace iterations with methods from potential theory. *SIAM Rev.*, 48:3–40, 2006.

[KV99] A. B. J. Kuijlaars and W. Van Assche. Extremal polynomials on discrete sets. *Proc. London Math. Soc.*, 79:191–221, 1999.

[Lag98] E. N. Laguerre. Le calcul des systèmes linéaires, extrait d'une lettre adressé à M. Hermite. In Ch. Hermite, H. Poincaré, and E. Rouché, editors, *Oeuvres de Laguerre*, volume 1, pages 221–267. Gauthier–Villars, Paris, France, 1898. The article is dated 1867 and is "Extrait du Journal de l'École Polytechnique, LXII^{e} Cahier".

[Lan50] C. Lanczos. An iteration method for the solution of the eigenvalue problem problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards*, 45:225–280, 1950.

[Lau77] T. C.-Y. Lau. Rational exponential approximation with real poles. *BIT*, 17:191–199, 1977.

[Lau87] D. P. Laurie. A recursive algorithm for the incomplete partial fraction decomposition. *Z. Angew. Math. Phys.*, 38:481–485, 1987.

[Leb77] V. I. Lebedev. On a Zolotarev problem in the method of alternating directions. *USSR Comput. Maths. Math. Phys.*, 17:58–76, 1977.

[LLL97] Y. Lai, K. Lin, and W. Lin. An inexact inverse iteration for large sparse eigenvalue problems. *Numer. Linear Algebra Appl.*, 4:425–437, 1997.

[LM98] R. B. Lehoucq and K. Meerbergen. Using generalized Cayley transformations within an inexact rational Krylov sequence method. *SIAM J. Matrix Anal. Appl.*, 20:131–148, 1998.

[LS94] A. L. Levin and E. B. Saff. Optimal ray sequences of rational functions connected with the Zolotarev problem. *Constr. Approx.*, 10:235–273, 1994.

[LS06] E. Levin and E. B. Saff. Potential theoretic tools in polynomial and rational approximation. In J.-D. Fournier et al., editors, *Harmonic Analysis and Rational Approximation*, volume 327 of *Lecture Notes in Control and Information Sciences*, pages 71–94. Springer-Verlag, Berlin, 2006.

[LSY98] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK User's Guide. Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia, PA, 1998.

[LT00] B. Le Bailly and J. P. Thiran. Optimal rational functions for the generalized Zolotarev problem in the complex plane. *SIAM J. Numer. Anal.*, 38:1409–1424, 2000.

[Mac46] C. C. MacDuffee. *The Theory of Matrices.* Chelsea Publishing, New York, 1946.

[Mag81] A. Magnus. Rate of convergence of sequences of Padé-type approximants and pole detection in the complex plane. In M. G. de Bruin and H. van Rossum, editors, *Padé Approximation and Its Applications*, volume 888 of *Lecture Notes in Mathematics*, pages 300–308. Springer-Verlag, Berlin, 1981.

[Mei64] G. Meinardus. *Approximation von Funktionen und ihre numerische Behandlung.* Springer-Verlag, Berlin, 1964.

[Mey00] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM, Philadelphia, PA, 2000.

[MN01a] I. Moret and P. Novati. An interpolatory approximation of the matrix exponential based on Faber polynomials. *J. Comput. Appl. Math.*, 131:361–380, 2001.

[MN01b] I. Moret and P. Novati. The computation of functions of matrices by truncated Faber series. *Numer. Funct. Anal. Optim.*, 22:697–719, 2001.

[MN04] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT*, 44:595–615, 2004.

[MN08] I. Moret and P. Novati. A rational Krylov method for solving time-periodic differential equations. *Appl. Numer. Math.*, 58:212–222, 2008.

[MNP84] A. McCurdy, K. C. Ng, and B. N. Parlett. Accurate computation of divided differences of the exponential function. *Math. Comp.*, 43:501–528, 1984.

[Mor07] I. Moret. On RD-rational Krylov approximations to the core-functions of exponential integrators. *Numer. Linear Algebra Appl.*, 14:445–457, 2007.

[Mor09] I. Moret. Rational Lanczos approximations to the matrix square root and related functions. *Numer. Linear Algebra Appl.*, 16:431–445, 2009.

[MS96] K. J. Maschhoff and D. C. Sorensen. P_ARPACK: An efficient portable large scale eigenvalue package for distributed memory parallel architectures. In J. Waśniewski et al., editors, *Applied Parallel Computing: Industrial Computation and Optimization*, volume 1184 of *Lecture Notes in Computer Science*, pages 478–486. Springer-Verlag, Berlin, 1996.

[MW05] B. V. Minchev and W. M. Wright. A review of exponential integrators for first order semi-linear problems. Preprint Numerics No. 2/2005, Norges Teknisk-Naturvitenskapelige Universitet, Trondheim, Norway, 2005.

[Neu98] H. Neuberger. Exactly massless quarks on the lattice. *Phys. Lett. B*, 417:141–144, 1998.

[Nev93] O. Nevanlinna. *Convergence of Iterations for Linear Equations*. Birkhäuser, 1993.

[NH95] I. Najfeld and T. F. Havel. Derivatives of the matrix exponential and their computation. *Adv. in Appl. Math.*, 16:321–375, 1995.

[NHA86] G. A. Newman, G. W. Hohmann, and W. L. Anderson. Transient electromagnetic response of a three-dimensional body in a layered earth. *Geophysics*, 51:1608–1627, 1986.

[Nør78] S. P. Nørsett. Restricted Padé approximations to the exponential function. *SIAM J. Numer. Anal.*, 15:1008–1029, 1978.

[Nou87] B. Nour-Omid. Lanczos method for heat conduction analysis. *Internat. J. Numer. Methods Engrg.*, 24:251–262, 1987.

[Nov03] P. Novati. Solving linear initial value problems by Faber polynomials. *Numer. Linear Algebra Appl.*, 10:247–270, 2003.

[NRT92] N. M. Nachtigal, L. Reichel, and L. N. Trefethen. A hybrid GMRES algorithm for nonsymmetric linear systems. *SIAM J. Matrix Anal. Appl.*, 13:796–825, 1992.

[NW77] S. P. Nørsett and A. Wolfbrandt. Attainable order of rational approximations to the exponential function with only real poles. *BIT*, 17:200–208, 1977.

[NW83] A. Nauts and R. E. Wyatt. New approach to many-state quantum dynamics: The recursive-residue-generation method. *Phys. Rev. Lett.*, 51:2238–2241, 1983.

[NW84] A. Nauts and R. E. Wyatt. Theory of laser-module interaction: The recursive-residue-generation method. *Phys. Rev.*, 30:872–883, 1984.

[OH84] M. L. Oristaglio and G. W. Hohmann. Diffusion of electromagnetic fields into a two-dimensional earth: A finite-difference approach. *Geophysics*, 49:870–894, 1984.

[Opi64] G. Opitz. Steigungsmatrizen. *Z. Angew. Math. Mech.*, 44:T52–T54, 1964.

[OR06] K. H. A. Olsson and A. Ruhe. Rational Krylov for eigenvalue computation and model order reduction. *BIT*, 46:S99–S111, 2006.

[Par88] O. G. Parfenov. Estimates of the singular numbers of the Carleson embedding operator. *Math. USSR Sb.*, 59:497–514, 1988.

[Par98] B. N. Parlett. *The Symmetric Eigenvalue Problem.* SIAM, Philadelphia, PA, 1998. Unabridged, corrected republication of 1980.

[PL86] T. J. Park and J. C. Light. Unitary quantum time evolution by iterative Lanczos reduction. *J. Chem. Phys.*, 85:5870–5876, 1986.

[Poi99] H. Poincaré. Sur les groupes continus. *Trans. Cambridge Phil. Soc.*, 18:220–255, 1899.

[PPT10] R. Pachón, R. B. Platte, and L. N. Trefethen. Piecewise smooth chebfuns. To appear in IMA J. Numer. Anal., 2010.

[PPV95] C. C. Paige, B. N. Parlett, and H. A. van der Vorst. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numer. Linear Algebra Appl.*, 2:115–133, 1995.

[Pro93] V. A. Prokhorov. Rational approximation of analytic functions. *Sb. Math.*, 184:3–32, 1993.

[Pro05] V. A. Prokhorov. On best rational approximation of analytic functions. *J. Approx. Theory*, 133:284–296, 2005.

[PS93] B. Philippe and R. B. Sidje. Transient solutions of Markov processes by Krylov subspaces. Research Report RR-1989, INRIA, Rennes, France, 1993.

[PS08] M. Popolizio and V. Simoncini. Acceleration techniques for approximating the matrix exponential operator. *SIAM J. Matrix Anal. Appl.*, 30:657–683, 2008.

[PT09] R. Pachón and L. N. Trefethen. Barycentric-Remez algorithms for best polynomial approximation in the chebfun system. *BIT*, 49:721–741, 2009.

[Rak96] E. A. Rakhmanov. Equilibrium measure and the distribution of zeros on the extremal polynomials of a discrete variable. *Sb. Math.*, 187:1213–1228, 1996.

[Ran95] T. Ransford. *Potential Theory in the Complex Plane.* Cambridge University Press, Cambridge, UK, 1995.

[Rei90] L. Reichel. Newton interpolation at Leja points. *BIT*, 30:332–346, 1990.

[Rin55] R. F. Rinehart. The equivalence of definitions of a matric function. *Amer. Math. Monthly*, 62:395–414, 1955.

[RS98] A. Ruhe and D. Skoogh. Rational Krylov algorithms for eigenvalue computation and model reduction. In G. Goos et al., editors, *Applied Parallel Computing*, volume 1541 of *Lecture Notes in Computer Science*, pages 491–502. Springer-Verlag, Berlin, 1998.

[Ruh83] A. Ruhe. Numerical aspects of Gram–Schmidt orthogonalization of vectors. *Linear Algebra Appl.*, 52/53:591–601, 1983.

[Ruh84] A. Ruhe. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra Appl.*, 58:391–405, 1984.

[Ruh94a] A. Ruhe. The rational Krylov algorithm for nonsymmetric eigenvalue problems III: Complex shifts for real matrices. *BIT*, 34:165–176, 1994.

[Ruh94b] A. Ruhe. Rational Krylov algorithms for nonsymmetric eigenvalue problems. *IMA Vol. Math. Appl.*, 60:149–164, 1994.

[Ruh94c] A. Ruhe. Rational Krylov algorithms for nonsymmetric eigenvalue problems. II: Matrix pairs. *Linear Algebra Appl.*, 197/198:283–296, 1994.

[Ruh98] A. Ruhe. Rational Krylov: A practical algorithm for large sparse nonsymmetric matrix pencils. *SIAM J. Sci. Comput.*, 19:1535–1551, 1998.

[Run84] C. Runge. Zur Theorie der eindeutigen analytischen Funktionen. *Acta Math.*, 6:229–244, 1884.

[SA00] W. D. Sharp and E. J. Allen. Stochastic neutron transport equations for rod and plane geometries. *Annals of Nuclear Energy*, 27:99–116, 2000.

[Saa92a] Y. Saad. Analysis of some Krylov subspace approximations to the exponential operator. *SIAM J. Numer. Anal.*, 29:209–228, 1992.

[Saa92b] Y. Saad. *Numerical Methods for Large Eigenvalue Problems.* Halsted Press, New York, 1992.

[Saa95] Y. Saad. Preconditioned Krylov subspace methods for the numerical solution of Markov chains. In W. J. Stewart, editor, *Computations with Markov Chains.* Kluwer Academic Publishers, Boston, MA, 1995.

[Sch90] M. J. Schaefer. A polynomial based iterative method for linear parabolic equations. *J. Comput. Appl. Math.*, 29:35–50, 1990.

[Sch07] T. Schmelzer. *The Fast Evaluation of Matrix Functions for Exponential Integrators.* PhD thesis, University of Oxford, Balliol College, Oxford, UK, 2007.

[Ser92] S. M. Serbin. A scheme for parallelizing certain algorithms for the linear inhomogeneous heat equation. *SIAM J. Sci. Stat. Comput.*, 13:449–458, 1992.

[SG04] O. Schenk and K. Gärtner. Solving unsymmetric sparse systems of linear equations with PARDISO. *Future Gener. Comp. Systems*, 20:475–487, 2004.

[SG06] O. Schenk and K. Gärtner. On fast factorization pivoting methods for sparse symmetric indefinite systems. *Electron. Trans. Numer. Anal.*, 23:158–179, 2006.

[SH82] R. M. Smith and A. G. Hutton. The numerical treatment of advection: A performance comparison of current methods. *Numer. Heat Transfer*, 5:439–461, 1982.

[Sim03] V. Simoncini. Restarted full orthogonalization method for shifted linear systems. *BIT*, 43:459–466, 2003.

[Sko96] D. Skoogh. *An Implementation of a Parallel Rational Krylov Algorithm.* PhD thesis, Chalmers University of Technology, Department of Computer Science, Göteborg, Sweden, 1996.

[Sko98] D. Skoogh. A parallel rational Krylov algorithm for eigenvalue computations. In G. Goos et al., editors, *Applied Parallel Computing*, volume 1541 of *Lecture Notes in Computer Science*, pages 521–526. Springer-Verlag, Berlin, 1998.

[SP99] P. Smit and M. H. C. Paardekooper. The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra Appl.*, 287:337–357, 1999.

[SS99] R. B. Sidje and W. J. Stewart. A numerical study of large sparse matrix exponentials arising in Markov chains. *Comput. Statist. Data Anal.*, 29:345–368, 1999.

[SS03] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM J. Sci. Comput.*, 25:454–477, 2003.

[SSV76] E. B. Saff, A. Schönhage, and R. S. Varga. Geometric convergence to e^{-z} by rational functions with real poles. *Numer. Math.*, 25:307–322, 1976.

[ST97] E. B. Saff and V. Totik. *Logarithmic Potentials with External Fields.* Springer-Verlag, Berlin, 1997.

[ST07a] T. Schmelzer and L. N. Trefethen. Computing the gamma function using contour integrals and rational approximations. *SIAM J. Numer. Anal.*, 45:558–571, 2007.

[ST07b] T. Schmelzer and L. N. Trefethen. Evaluating matrix functions for exponential integrators via Carathéodory–Fejér approximation and contour integrals. *Electron. Trans. Numer. Anal.*, 29:1–18, 2007.

[Sta89] H. Stahl. General convergence results for rational approximants. In C. K. Chui, L. L. Schumaker, and J. D. Ward, editors, *Approximation Theory VI*, pages 605–634. Academic Press, Boston, MA, 1989.

[Sta91] G. Starke. Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 28:1431–1445, 1991.

[Sta96] H. Stahl. Convergence of rational interpolants. *Bull. Belg. Math. Soc. Simon Stevin*, 3:11–32, 1996.

[Sta03] H. Stahl. Best uniform rational approximation of x^α on $[0, 1]$. *Acta Math.*, 190:241–306, 2003.

[Ste93] F. Stenger. *Numerical Methods Based on Sinc and Analytic Functions.* Springer-Verlag, Berlin, 1993.

[Ste94] F. Stenger. Numerical methods via transformations. In R. V. M. Zahar, editor, *Approximation and Computation: A Festschrift in Honor of Walter Gautschi.* Birkhäuser, Boston, MA, 1994.

[Ste98] G. W. Stewart. *Afternotes Goes to Graduate School. Lectures on Advanced Numerical Analysis.* SIAM, Philadelphia, PA, 1998.

[Ste00] F. Stenger. Summary of Sinc numerical methods. *J. Comput. Appl. Math.*, 121:379–420, 2000.

[Sti81] L. Stickelberger. Zur Theorie der linearen Differentialgleichungen. Teubner, Leipzig, 1881. Akademische Antrittsschrift.

[SV00] Y. Saad and H. A. van der Vorst. Iterative solution of linear systems in the 20th century. *J. Comput. Appl. Math.*, 123:1–33, 2000.

[Syl83] J. J. Sylvester. On the equation to the secular inequalities in the planetary theory. *Phil. Mag.*, 16:267–269, 1883.

[Tal79] A. Talbot. The accurate numerical inversion of Laplace transforms. *J. Inst. Maths. Applics.*, 23:97–120, 1979.

[Tal91] H. Tal-Ezer. High degree polynomial interpolation in Newton form. *SIAM J. Sci. Stat. Comput.*, 12:648–667, 1991.

[Tay43] A. E. Taylor. Analysis in complex Banach spaces. *Bull. Amer. Math. Soc.*, 49:652–669, 1943.

[Tay50] A. E. Taylor. Spectral theory for closed distributive operators. *Acta Math.*, 84:189–224, 1950.

[TB97] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra.* SIAM, Philadelphia, PA, 1997.

[TG83] L. N. Trefethen and M. H. Gutknecht. The Carathéodory–Fejér method for real rational approximation. *SIAM J. Numer. Anal.*, 20:420–436, 1983.

[THP+09] L. N. Trefethen, N. Hale, R. B. Platte, T. A. Driscoll, and R. Pachón. Chebfun Version 3. Oxford University, 2009. `http://www.maths.ox.ac.uk/chebfun/`.

[Tod84] J. Todd. Applications of transformation theory: A legacy from Zolotarev (1847–1878). In S. P. Singh, editor, *Approximation Theory and Spline Functions*, pages 207–245. D. Reidel Publishing, Dordrecht, Netherlands, 1984.

[Tre81] L. N. Trefethen. Rational Chebyshev approximation on the unit disk. *Numer. Math.*, 37:297–320, 1981.

[Tre83] L. N. Trefethen. Chebyshev approximation on the unit disk. In H. Werner et al., editors, *Computational Aspects of Complex Analysis*, pages 309–323. D. Reidel Publishing, Dordrecht, Netherlands, 1983.

[TT07] J. D. Tebbens and M. Tůma. Efficient preconditioning of sequences of nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 29:1918–1941, 2007.

[TWS06] L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer. Talbot quadratures and rational approximations. *BIT*, 46:653–670, 2006.

[Tyr96] E. Tyrtyshnikov. Mosaic-skeleton approximations. *Calcolo*, 33:47–57, 1996.

[Van77] C. F. Van Loan. The sensitivity of the matrix exponential. *SIAM J. Numer. Anal.*, 14:971–981, 1977.

[Vor65] Y. V. Vorobyev. *Method of Moments in Applied Mathematics*. Gordon and Breach Science Publishers, New York, 1965.

[Vor87] H. A. van der Vorst. An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A. *J. Comput. Appl. Math.*, 18:249–263, 1987.

[Wal65] J. L. Walsh. Hyperbolic capacity and interpolating rational functions. *Duke Math. J.*, 32:369–379, 1965.

[Wal69] J. L. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Domain*. AMS, Providence, RI, 5th edition, 1969.

[WT07] J. A. C. Weideman and L. N. Trefethen. Parabolic and hyperbolic contours for computing the Bromwich integral. *Math. Comp.*, 76:1341–1356, 2007.

[Zol77] E. I. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Zap. Imp. Akad. Nauk St. Petersburg*, 30:1–59, 1877. In Russian.